TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

CHUN WANG

DESIGN AND ANALYSIS OF CODED APERTURE FOR 3D SCENE SENSING

Master of Science thesis

Examiner: Prof. Atanas Gotchev
Examiner: Prof. Ulla Ruotsalainen
Examiner and topic approved by the
Faculty Council of the Faculty of
Natural Sciences
on 7th May 2014

ABSTRACT

CHUN WANG: Design and analysis of coded aperture for 3D scene sensing
Tampere University of Technology
Master of Science thesis, 69 pages, 0 Appendix pages
January 2015
Master's Degree Programme in Biomedical Engineering
Major: Medical Informatics
Examiner: Prof. Atanas Gotchev
Examiner: Prof. Ulla Ruotsalainen
Keywords: defocus blur, depth from defocus, inverse problem, depth estimation, coded aperture

In this thesis, the application of coded aperture in depth estimation is studied. More specifically, depth from defocus (DfD) is considered. DfD is a popular computer vision technique, which utilises the defocus blur cue for depth estimation. A general review of studies about the defocus blur, both its properties as a depth cue and its relation with the disparity cue, is presented. DfD methods are comprehensively investigated under two types of solving strategies. One is image restoration-based, whose success depends on the quality of image restoration; while the other strategy directly focuses on the depth estimation without requiring image restoration, and thus is referred to as the restoration-free strategy. The defocus blur is actually characterised by the point spread function (PSF) of the camera imaging system. The PSF of the camera can be modified by inserting a physical mask in the camera aperture position. A recent technique called coded aperture, which refers to the insertion of a coded mask in the aperture position, utilises this fact to improve the performance of DfD. Optimisation of the mask pattern for depth estimation is discussed in detail. A camera with a coded mask is built. The existing coded aperture methods for depth estimation are implemented and tested in both simulations and real experiments. Wave-optics based PSF calculation is proposed to have an accurate imaging model and avoid capturing PSFs in real experiments.

Finally, several stereo cameras equipped with different sets of masks are analysed to explore the possible improvements in depth estimation by jointly utilising disparity and defocus blur cues. Results show that DfD can give valuable complementary depth information to stereo vision where stereo matching suffers from the correspondence problem. On the other hand, a stereo camera arrangement is shown to be useful for getting a single shot coded aperture system which employs a pair of complementary masks. A modified DfD algorithm is developed for that system.

PREFACE

This thesis work is done with the 3D media group in the Department of Signal Processing (SGN) at Tampere University of Technology (TUT). The first aim is to understand and study coded aperture for depth estimation and demonstrate the understanding with simulations and experiments. The second aim is to explore the possibility of combining coded aperture and stereo matching.

I gratefully thank my thesis supervisor Prof. Atanas Gotchev, for his continuous support and guiding. I also thank all members of the 3D group and others, for all kinds of help and suggestions, especially Dr. Atanas Boev, Ahmed Durmush, Dr. Erdem Sahin, Mihail Georgiev, Olli Suominen and Dr. Suren Vagharshakyan. I could not finish this thesis without benefiting from their profound knowledge and skills. Special thanks to Dr. Suren Vagharshakyan for lending me the most impressive inverse problem book, which helped quite a lot, and to Dr. Robert Bregovic for his radiant optimism.

I also would like to express my appreciation to all professors and lecturers who taught me all kind of knowledge, which give me confidence to explore this unknown area, a small one, though.

Last but not least, I thank the department secretaries Susanna Anttila and Virve Larmila from SGN, and coordinator Ulla Siltaloppi from the international office at TUT, who helped me to get familiar with the working environment and handle all administrative matters.

Tampere, 2.12.2014

Chun Wang

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND SYMBOLS

2D	Two-dimensional
3D	Three-dimensional
AMA	Accuracy maximising analysis
CoC	Circle of confusion
DfD	Depth from defocus
DFT	Discrete Fourier transforms
IRLS	Iterative re-weighted least squares
LCA	Liquid crystal array
LCoS	Liquid crystal on silicon
MAP	Maximum *a posteriori*
MLE	Maximum likelihood estimation
MRF	Markov random field
NSR	Noise-to-signal ratio
OTF	Optical transfer function
PSF	Point spread function
SNR	Signal-to-noise ratio
SVD	Singular value decomposition
TVR	Threshold versus reference

$\boldsymbol{a}$	Transmission efficiency
$\mathcal{A}$	An operator representing the role of imaging system
B	Baseline width
$\mathcal{B}$	Frequency support of a PSF
$\boldsymbol{c}$	camera settings/parameters
$\boldsymbol{C_N}$	A vector of N points' camera settings
d	Depth
d_f	Focused distance
$disp$	Disparity
d_L	Lens aperture diameter
$\boldsymbol{D_N}$	A vector of N points' depths, or depth map
$\mathcal{D}_{\mathbb{R}^2}$	A sub-domain of the continuous scene plane
$\mathcal{D}_\Gamma$	A sub-domain of the continuous image plane
$\boldsymbol{Disp_M}$	Depth map in disparity values
f	Focal length
f^0	Continuous scene intensity function
$\boldsymbol{f_N^0}$	Scene intensity vector

F^0	Fourier transform of f^0
$\boldsymbol{f_M}$	All-in-focus image
$\mathcal{F}$	A filter bank
g	Continuous noisy image
g^0	Continuous noise free image
$\boldsymbol{g_M}$	Noisy image vector
G^0	Fourier transform of g^0
$\boldsymbol{h}^{\boldsymbol{c},d}$	Discrete point spread function
$\boldsymbol{H}_{\boldsymbol{c},d}$	Camera system matrix
$\boldsymbol{H}^{\boldsymbol{c},d}$	Discrete Fourier transform of $\boldsymbol{h}^{\boldsymbol{c},d}$
$\boldsymbol{H}^{\perp}_{\boldsymbol{c},d}$	Operator projecting to the orthogonal subspace
$k^{\boldsymbol{c},d}$	Continuous point spread function
K_{max}	An upper bound of k
$K^{\boldsymbol{c},d}$	Fourier transform of k
$\mathcal{K}$	A set of depths
l_f	Distance between the lens and the image plane
$\boldsymbol{L_M}$	A sub-domain of the discrete image plane
$\boldsymbol{L_N}$	A sub-domain of the discrete scene plane
$M(\boldsymbol{\eta})$	Mask function
N_{pix}	Number of pixels
$\boldsymbol{p}$	Vector tracing the scene
$p_{\boldsymbol{m}}$	A weight kernel representing the detector's response
$\mathcal{P}_{\mathcal{B}}$	Band-limiting operator
$\boldsymbol{Q}$	Information other than the PSF
$\boldsymbol{\mathcal{R}}$	Features
$\mathbb{R}$	The set of real numbers
s_{pix}	Pixel pitch
S_{coc}	Physical size of the circle of confusion
$\boldsymbol{\mathcal{X}}$	Scene space
$\boldsymbol{\mathcal{X}_N}$	Scene intensity vector space
$\boldsymbol{\mathcal{Y}}$	Image space
$\boldsymbol{\mathcal{Y}_M}$	Image vector space
$\mathbb{Z}^+$	The set of positive integers
α	Lens magnification
Γ	Image plane
λ	Wavelength
Λ	Spectral components
$\boldsymbol{\Psi}$	Linear weight matrix in the frequency domain
ω	Continuous sensor noise

$\boldsymbol{\omega_M}$	Sensor noise vector
∇	Derivative operator
$\bullet$	Element-wise multiplication
$\otimes$	Convolution
$\| \|_p$	p-norm
$\| \|^2$	Element-wise square

1. INTRODUCTION

Depth perception, which is defined as the ability to extract three-dimensional (3D) representations of physical reality from two-dimensional (2D) retinal images, is a born gift to the human being. With the ability to judge the distance, we can locate an object in space and estimate its size. This ability is essential for our survival since most of the activities like jumping and grasping cannot be achieved without it. Nowadays the depth information is not only needed for the daily life of a human being, but also needed in many engineering fields like multimedia and computer vision. Since the development of the vision related technologies are ever increasing, inferring depth from images and videos becomes demanding and forms a base of many fascinating areas, e.g. virtual reality and robot navigation. However, what cameras record are 2D images that are results of projection of the 3D world, so it is not a trivial task to infer the (correct) depth from them.

Depth perception in human vision and depth estimation in computer vision have both common and different properties. In human vision, it has been shown that there are several factors related with depth information, referred to as depth cues, playing key roles in the depth inferring process in the brain. In computer vision, similar is true, and most of the depth cues are also available. In human vision, where the mysterious brain can utilise all depth cues simultaneously to interpret the 3D world automatically, many people can benefit from it without even being aware of it, let alone understanding the mechanism behind it. In computer vision, however, the situation varies with the chosen depth cue, the technique and the algorithm. Indeed, developing techniques and algorithms to utilise certain depth cues are the main issues for depth estimation in computer vision [41].

This thesis is aimed at studying techniques and algorithms that mainly utilise the defocus blur cue for depth estimation. As a relatively new depth cue, the defocus blur cue gains growing popularity in computer vision. The most popular technique utilising the defocus blur cue to infer depth is known as depth from defocus (DfD) in the literature, which includes a class of implementations with varying settings and/or algorithms. Among those implementations, recently a branch of DfD techniques utilising coded aperture is of particular interest. In this branch of DfD techniques,

instead of conventional cameras, cameras equipped with a mask in the aperture position are employed to sense the 3D world. By utilising masks of different patterns, a coded aperture camera can cause different defocus blurring effects, and some of those effects may improve the depth estimation result. In addition to studying the defocus blur cue alone, it is also interesting to exploit its relationship with the disparity cue, which is a well-known depth cue and has been widely used in computer vision.

The properties of the defocus blur cue and its relation with the disparity cue are investigated in Chapter 2. In Chapter 3, the camera imaging system is modelled. Then two strategies for solving DfD are introduced in Chapter 4. The principle of coded aperture and mask design are reviewed in Chapter 5. Simulation and experimental results of coded aperture are presented and discussed in Chapter 6. In Chapter 7, the possibility of using the disparity cue and the defocus blur cue in combination is explored and two types of coded aperture stereo camera systems are proposed.

2. DISPARITY CUE AND DEFOCUS BLUR CUE

In this chapter, two depth cues, the disparity cue and the defocus blur cue, are studied. Unlike the disparity cue, which has long been known and well analysed, the defocus blur cue, which is going to be used intensively in the following chapters, is relatively new, and thus more efforts are paid on understanding its properties as a depth cue. Particularly, it is also interesting to compare those two depth cues and to explore the possibility of using them jointly.

2.1 Disparity cue

The disparity cue is a primary cue in human vision, and it is also the most popular depth cue in computer vision. Since it has been extensively studied, here we just include the relevant information necessary for other sections, for more information please refer to [47].

As a binocular cue, the disparity cue is encoded in two views. In human vision, it is defined as the location difference of the same object between its projections on the left and the right eyes. This location difference is known as the retinal disparity and is a result of the fact that two eyes see from slightly different positions. The retinal disparity of a point reflects its depth related to the fixation point. As shown in Figure 2.1, for a fixation point, it projects on the same positions on both eyes and thus cause no retinal disparity; while for the point deviating from the fixation point, the magnitude of retinal disparity reflects its relative depth to the fixation point and the orientation of retinal disparity indicates the side of the point related to the fixation point. However, when a point deviates from the fixation point too much, its depth cannot be inferred from the retinal disparity. That is, the retinal disparity cue has a limited working range, which is reported by Schor and Wood to be within roughly 0.25 - 40 arc min [44].

In computer vision, two eyes are replaced with two cameras. However, unlike eyes that fixate on a particular location, two cameras are usually put in parallel, and this arrangement is referred as the stereo camera setup, where the distance between two cameras is called the baseline B, as shown in Figure 2.2. By using triangulation,

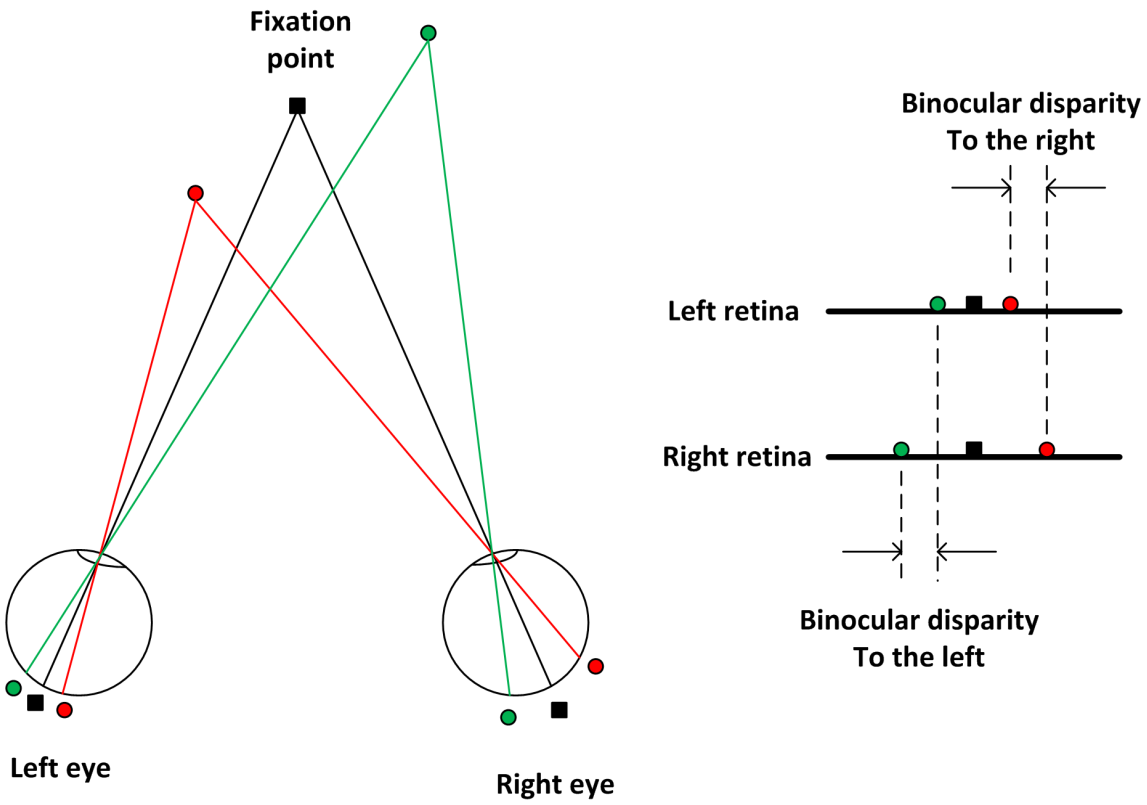


Figure 2.1 *Illustration of the disparity in human vision (adapted from Figure 1 in [37]).*

we can derive the relation between depth d and disparity $disp$ as

$$disp = \frac{fB}{d}, \tag{2.1}$$

where f is the focal length, corresponding to the distance between the lens and the image plane in the pinhole camera model. This relation reveals that under the stereo camera setup, the disparity is inversely proportional to the depth, as shown in Figure 2.3. If the same discrimination criteria apply to the whole depth range, the depth resolution provided by the disparity cue decreases as the depth increases. As a consequence of this relation, the disparity cue in computer vision also has a working range.

2.2 Defocus blur cue

In contrast with the disparity cue, the defocus blur cue is a monocular pictorial cue. It is widely known that most of biological and artificial lens systems can only bring objects close to the focused distance into focus. Therefore, when a 3D scene is recorded in 2D images, it is inevitable to see that objects at other distances

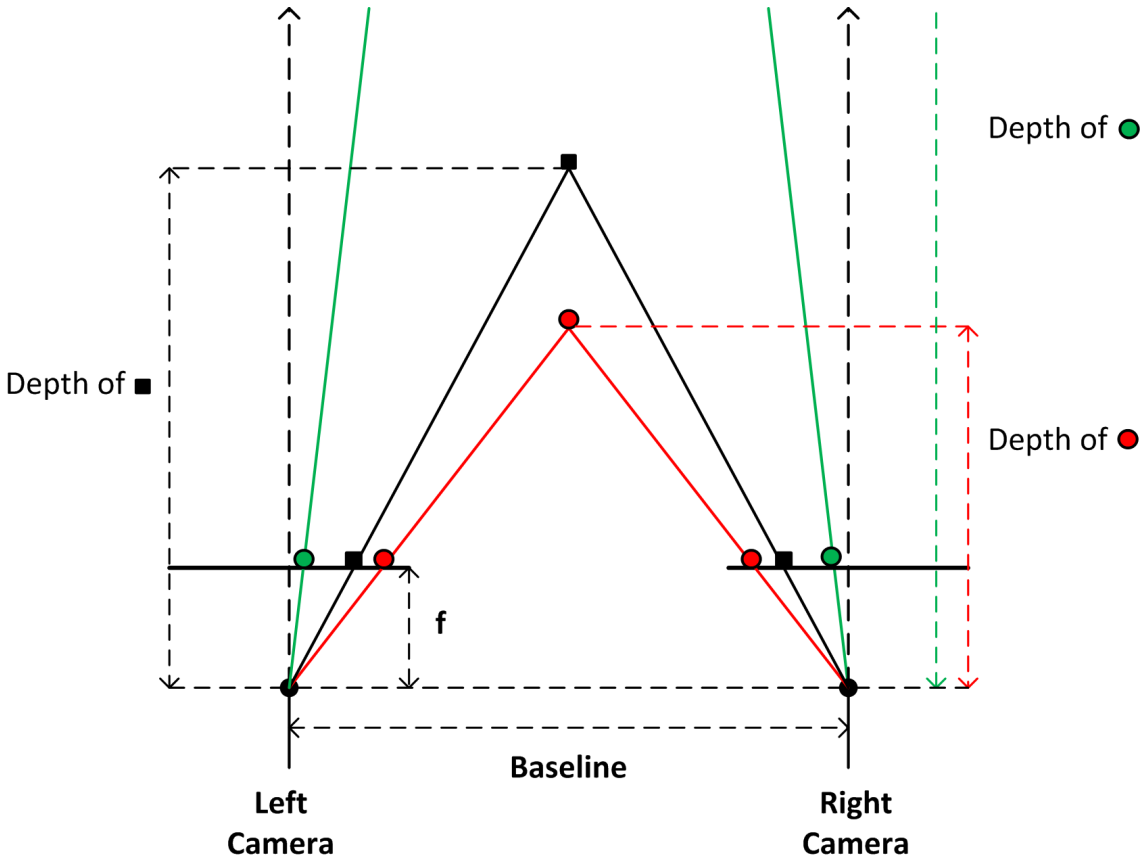


Figure 2.2 *Illustration of the disparity in computer vision.*

are blurred in images. That is, most optical systems have limited depth of field. Generally, this phenomenon is unfavoured and is treated as a drawback of the optical system. However, Pentland [36] pointed out that the degree of blur can reflect the depth between the object and the focused distance; therefore, it can actually serve as a depth cue.

In human vision, Marshall et al. [28] and Mather [31] independently conducted similar experiments and reported that the degree of blur at the boundary between a focused surface and a defocused surface is important, and it may be used to determine depth orders. For example, as illustrated in Figure 2.4(a), the surface having the same state as the boundary is seen as nearer and occluding the other. In addition, Mather [31] showed that besides the boundary blur, the region blur can also enhance depth perception. An example is shown in Figure 2.4(b), and it shows that when the background is blurred, it can enhance a feeling that the sharp central square is floating above it. Furthermore, Mather and Smith [34] studied the effectiveness of the boundary blur discrimination and region blur discrimination affecting depth ordering, but they reported that the boundary blur acts as a depth cue only when it is either not blurred or extremely blurred, and it may indicate that

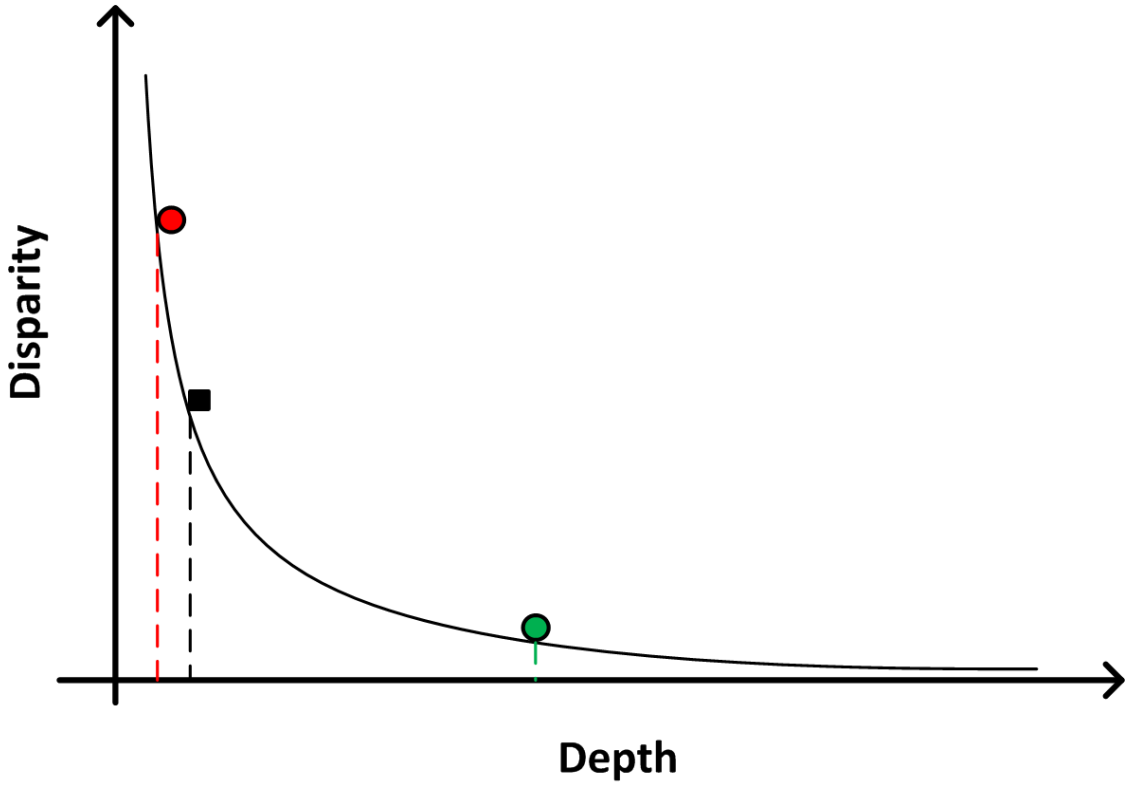
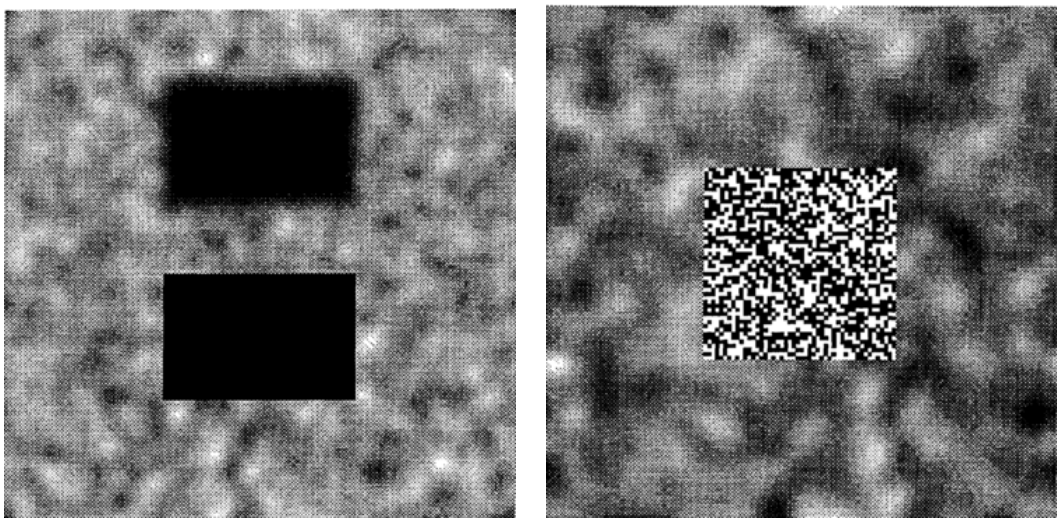


Figure 2.3 *Illustration of the depth-disparity relation in computer vision.*



(a) An example of the defocus blur degree at the boundary affecting depth ordering.

(b) An example of the defocus blur degree of an area affecting depth sensing.

Figure 2.4 *Examples of the defocus blur cue affecting depth perception in human vision [32]. Reprinted by permission of Pion Ltd, London, www.pion.co.uk and www.envplan.com*

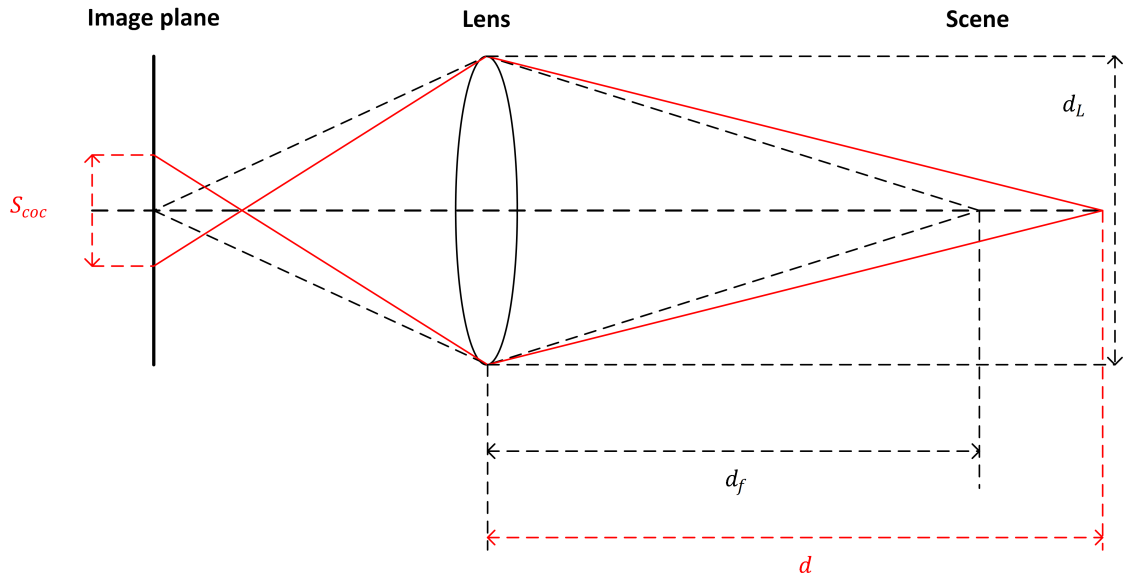


Figure 2.5 *Illustration of the defocus blur cue with thin-lens camera model.*

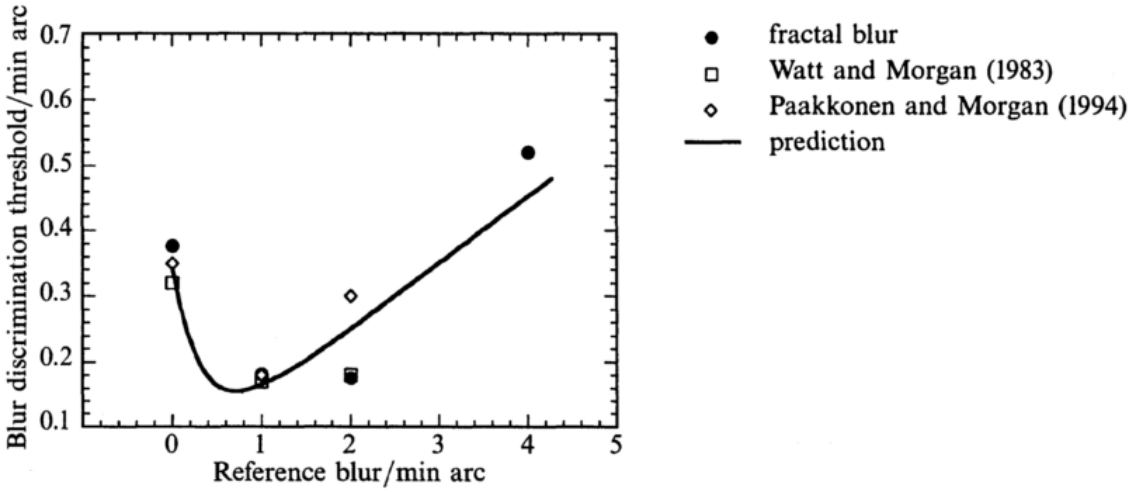


Figure 2.6 *The blur discrimination thresholds in human vision [32]. Reprinted by permission of Pion Ltd, London, www.pion.co.uk and www.envplan.com*

the defocus blur cue is an insignificant depth cue.

Since the degree of defocus blur variance is relatively insignificant, it becomes an important question that how sensitive the vision system is to the small degree of defocus blur variance. In human vision, a series of studies were done to determine the blur threshold and the blur discrimination for blur detection. Their results are consistent and show that the defocus blur detection threshold is roughly 0.4 - 1 arc min, and the blur discrimination threshold is related to the reference blur.

This relation is best viewed in the threshold versus reference (TVR) curve. One result reported by [32] is shown in Figure 2.6. When the reference blur is small(<1 arc min), the blur discrimination threshold decreases as the reference blur increases; after that, it increases with the increase in the reference blur accordingly. As pointed out by Mather [32], the TVR curve indicates that the human vision system is unable to use the defocus blur as a depth cue within the range just around to the fixation point. For a complete review, please refer to [56]. In conclusion, in human vision, due to the poor blur discrimination ability, the defocus blur cue should be viewed as a qualitative cue [34], [54].

In computer vision, the quantitative analyses can be conducted to understand the physical properties of the defocus blur cue. By utilising the thin-lens camera model and the geometrical optics, the relation between the depth d and the degree of defocus blur, characterised by the size S_{coc} of circle of confusion (CoC), is as follows:

$$S_{coc} = d_L \left(\left| \frac{f d_f}{(d_f - f)d} - \frac{d_f}{d_f - f} + 1 \right| \right) \tag{2.2}$$

$$\approx \frac{f d_L}{d}, \tag{2.3}$$

where f is the focal length of the lens, d_L is the diameter of lens aperture and d_f is the focused distance, as denoted in Figure 2.5. Please notice that when $d_f \gg f$, the depth-defocus blur degree relation is independent to the focused distance, as shown in Eq. (2.3). Nevertheless, the blur discrimination ability depends on the quality of the optical system and the method used to detect the degree of defocus blur.

2.3 Relation between disparity cue and defocus blur cue

Studying the relation between the two depth cues mentioned above is an interesting and important topic. Since the very beginning, the defocus blur cue has been compared to the disparity cue, which is a primary cue in human vision as well as the most popular depth cue in the field of computer vision.

In computer vision, based on the analyses done in [43], two depth cues share the same principle but differ in scales, and what leads to this scale difference is the physical size of the lens aperture diameter in the case of the defocus blur cue, or the baseline width in the case of the disparity cue, as can be learnt from Eq. (2.1) and Eq. (2.3). Since the defocus blur cue is a monocular cue, its scale is constrained by the lens aperture diameter, which in fact plays the role of baseline in the case of depth-disparity relation. The depth resolution provided by the disparity cue is better than that provided by the defocus blur cue, since in most practical applications

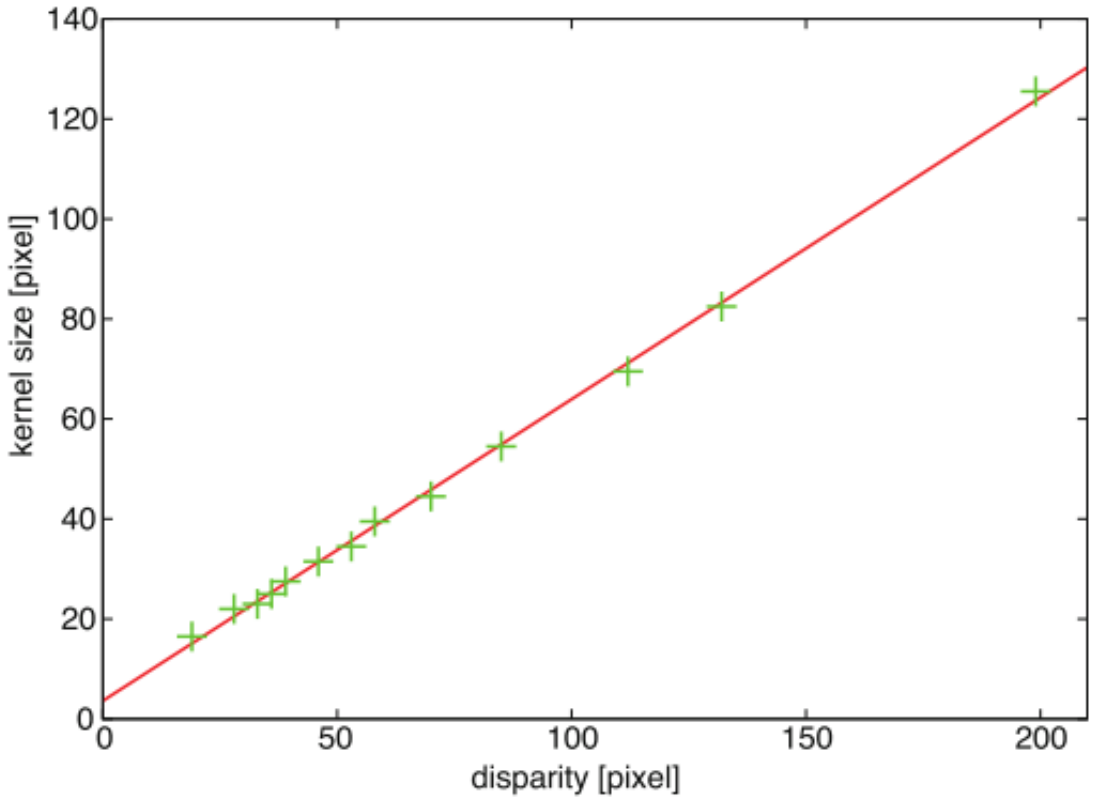


Figure 2.7 *An example of disparity-defocus blur degree relation [51]. Reprinted by permission. ©2013 IEEE*

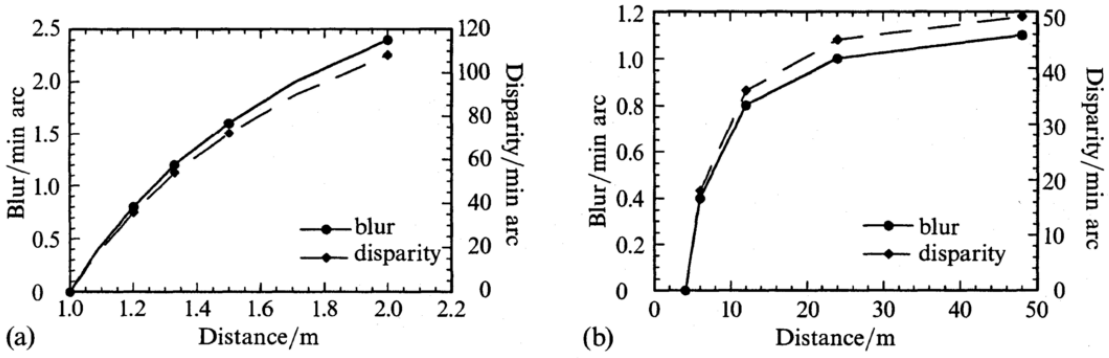


Figure 2.8 *The depth-retinal disparity relation (broken lines) and the depth-defocus blur degree relation. Left: Fixation at 1 metre. Right: Fixation at 4 metres [32]. Reprinted by permission of Pion Ltd, London, www.pion.co.uk and www.envplan.com*

of computer vision, the baseline is wider compared to the lens aperture diameter. That is, for the same amount of depth variance, the disparity value variance is more significant than the variance of defocus blur degree. Although according to Eq. (2.3), using a lens with larger aperture diameter and longer focal length can result in more significant variances, they are still relatively less significant than the disparity variance. It has also been shown experimentally that the two depth cues perform in the same way, besides the scale. One example is given in Figure 2.7, where Takeda et al. [50], [51] experimentally showed that when two cameras are almost focused on the infinity, the relation between the disparity and the degree of defocus blur is linear, and the slope can be inferred as the ratio between lens aperture diameter and baseline.

In human vision studies, similar opinion is adopted. By using Eq. (2.1) and Eq. (2.3), Mather [32] showed that the disparity cue is more significant than the defocus blur cue, as shown in Figure 2.8. Regarding the discrimination ability, researchers found that a small disparity variance is more detectable than a small variance in the degree of defocus blur. That is, the variance of defocus blur degree needs to be sufficiently large to be noticed, due to the poor blur discrimination ability [34].

2.4 Interaction between disparity cue and defocus blur cue

The human visual system uses several depth cues to infer depth information, how do those different information sources interact with each other? In this section, this important question is narrowed down to the interaction between the defocus blur cue and the disparity cue.

In human vision studies, based on the curve in Figure 2.8, together with the disparity covering range and the blur detection threshold, given in Section 2.1 and Section 2.2 respectively, Mather [32] suggested that the disparity cue and the defocus blur cue may serve in different depth ranges. His further studies with Smith [33] support this suggestion by noticing that within the valid range of disparity cue, image blur has insignificant effects on it. Therefore, it is more likely that the disparity cue is used for distances near the fixation point while the defocus blur cue takes over in longer distances. This complementary relation is also confirmed by other researchers, e.g. [40], [16].

In computer vision, the idea of combining those two depth cues has also gained popularity, in order to increase depth estimation results' quality [39], [42], [14], [51], [52]. The motivations behind those studies mainly are based on two differences. One is that those two depth cues respond to the same amount of depth variance in different

	Implementation	Occlusions	Repeating Patterns	Bright/Dark Features	Noise
Defocus	+ no calibration needed - aperture-size dependent - patch size dependent	+ easily affected - more stable	+ contrast detection distinguishes	- contrast detection ambiguous	+ 2D blur kernel provides better support with noise
Correspondence	+ not dependent on DOF - noise from using pinhole - correspondence problem	+ less affected - unstable if affected	- correspondence ambiguity	+ correspondence not affected as much	- matching prone to noise - pinhole image noise

Figure 2.9 *A comparison of using the defocus blur cue and the disparity cue [52]. Reprinted by permission. ©2013 IEEE*

scales. As a monocular cue, the defocus blur cue is less affected by problems like occlusions, which are known to be painful for the disparity cue. The other is that the methods used to extract those two depth cues are different. The disparity cue is extracted by finding the correspondence in different views, which fails in regions, e.g. with repetitive patterns or edges along the epipolar line; while the defocus blur cue is extracted by a comparison between images captured from the same view, and thus is stable to repetitive patterns. Those two differences may lead to a complementary performance of two cues, as summarised in Figrue 2.9. In computer vision, the defocus blur cue is used in the same depth range as the disparity cue, which differs with the human vision case, where those two depth cues are shown to complement each other in covering complementary depth ranges.

3. CAMERA IMAGING SYSTEM

The camera imaging system is responsible for image capture and processing from image formation to storage. Its understanding is essential for interpreting the depth. This chapter addresses the problem of modeling the camera imaging system. As pointed out in [4], an image is a degraded representation of the original 3D scene, and the degradation is mainly introduced during the image formation process and the recording process, denoted by blurring and noise, respectively. Among multiple reasons causing blurring, here only the blurring caused by the defocus is considered for the problem of depth estimation via defocus blur cue. Therefore, during the discussion below, both the camera and the scene are assumed to be perfectly fixed, which eliminates the influence of motion blurring. Also, the lens is assumed to be free of aberrations.

3.1 Space variant imaging system

In order to describe a camera imaging system, three parts are needed: a 3D scene to be imaged as the signal source, a camera imaging system that captures and processes the signal and the captured images as the result of this processing.

Firstly, the 3D scene is considered. In most cases, a 3D scene can be viewed as a cloud of self-luminous point light sources representing all the visible parts of objects in this 3D scene. For each point light source, its position on the scene space can be traced by a vector $\boldsymbol{p}$, and $\boldsymbol{p} \in \mathbb{R}^3$. That is, the vector $\boldsymbol{p}$ traces the surface of objects in the 3D scene. This vector $\boldsymbol{p}$ can be further separated into two parts, one part is $\boldsymbol{p}_{\boldsymbol{x}} = [\boldsymbol{p}_{x_1}, \boldsymbol{p}_{x_2}]^\top \in \mathbb{R}^2$ denoting the position on the scene plane, the other is $\boldsymbol{p}_d \in \mathbb{R}$ denoting the depth. That is, $\boldsymbol{p} = [\boldsymbol{p}_{\boldsymbol{x}}, \boldsymbol{p}_d]^\top$. One point light source is shown in Figure 3.1 as an example.

According to [4], under the Lambertian assumption, the appearance of a 3D scene can be considered as an unknown spatial intensity distribution over the space and denoted by $f^0(\boldsymbol{p})$, which is therefore known as the scene intensity function. Particularly, in most of the cases, a scene intensity function contains finite energy, that

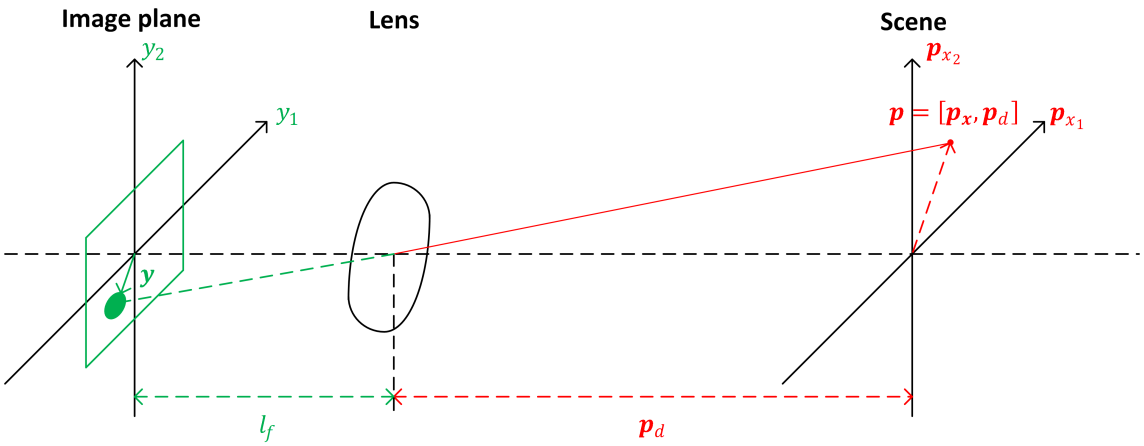


Figure 3.1 *Illustration of the image formation process and the coordinate system, where the lens centre is taken as the origin.*

is,

$$\int_{\mathbb{R}^3} \left| f^0(\boldsymbol{p}) \right|^2 \mathrm{d}\boldsymbol{p} < \infty. \tag{3.1}$$

It means that scene intensity functions are square integrable and thus form a $\mathcal{L}^2(\mathbb{R}^3)$-space, which is known as the scene space and is denoted by $\boldsymbol{\mathcal{X}}$. Since a $\mathcal{L}^2(\mathbb{R}^3)$-space is also a Hilbert-space, the scene space $\boldsymbol{\mathcal{X}}$ is naturally equipped with the inner product as follows

$$(f_1, f_2) = \int_{\mathbb{R}^3} f_1(\boldsymbol{p}) \bar{f}_2(\boldsymbol{p}) \mathrm{d}\boldsymbol{p}, \tag{3.2}$$

where $\bar{f}_2$ represents the complex conjugate of f_2.

Secondly, how the camera imaging system transforms the signals from the scene space to the image plane is studied. In general, the role of imaging system can be treated as an operator, denoted by $\mathcal{A}$, which maps a scene intensity function $f^0(\boldsymbol{p})$ of $\boldsymbol{\mathcal{X}}$ to its noise free image $g^0(\boldsymbol{y})$, as follows

$$g^0 = \mathcal{A} f^0. \tag{3.3}$$

Specifically, in the case of camera imaging system, the operator $\mathcal{A}$ can be replaced

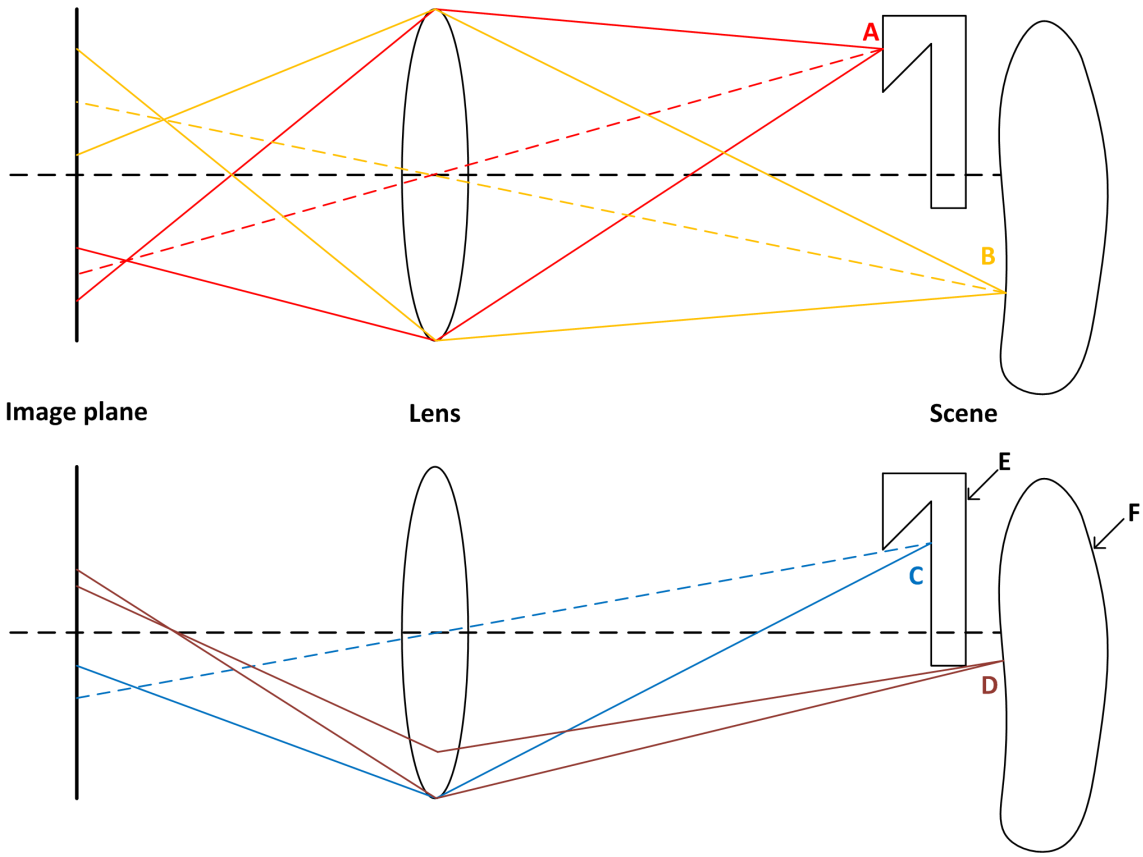


Figure 3.2 *Illustration of point light sources of three categories.*

by an integral operator as follows

$$g^0(\boldsymbol{y}) = \int_{\mathbb{R}^3} k\left(\boldsymbol{y}, \boldsymbol{p}\right) f^0(\boldsymbol{p}) \mathrm{d}\boldsymbol{p}, \tag{3.4}$$

where $k\left(\boldsymbol{y}, \boldsymbol{p}\right)$ is known as the point spread function (PSF) or the impulse response of the system [4].

In a camera imaging system, a PSF $k\left(\boldsymbol{y}, \boldsymbol{p}\right)$ is known as the image of an unit intensity point light source $\boldsymbol{p}$ in the image plane, as shown in Figure 3.1. Consequently, in Eq. (3.4), g^0 is actually modelled as a superposition of images of all points of f^0. In addition, since it is the PSF that causes the blurring effect, g^0 is also known as a blurred image of the corresponding scene f^0 [4].

There are several factors that can affect a PSF, and one of them of interest here is the defocus, or equivalently, out-of-focus. As shown in Figure 2.5, a point deviating from the focused distance on the scene results in a small area in the image plane, which is known as the CoC, inside which the intensity is assumed to be nearly uniform according to the geometrical optics. However, for a more rigorous treat-

ment, the diffraction effects should be taken into account, as will be discussed in Section 6.1. According to the thin lens model, the camera setting parameters are mainly the aperture shape, the focal length and the focused distance. For capturing a still image, all those parameters, together with the camera's position and viewing direction, are fixed, so it can be assumed that they are all well set and denoted by $\boldsymbol{c}$. However, due to the limited physical size and viewing angle of a lens as well as the complex structure of a 3D scene, there generally exist occlusions between different objects in the 3D scene and/or self-occlusions between different parts of the same object. Consequently, not all point light sources of the scene are equally visible by the lens. As illustrated in Figure 3.2, point light sources form three categories. Point light sources of the first category are not occluded and thus the whole lens 'sees' them, like point light sources A and B. Those belonging to the second category are partially occluded, like point light sources C and D. For this case, parts of the lens 'sees' those points while the rest parts do not. Finally, the point light sources belonging to the third category are totally occluded and thus are invisible to the lens, like E and F. In order to deal with this issue, the concept of the effective aperture shape is introduced. For each point light source in the scene, the visible part of the aperture is described. Obviously, the effective aperture shape varies over point light sources. Since the effective aperture shape can be considered as a part of the camera setting $\boldsymbol{c}$, the camera setting $\boldsymbol{c}(\boldsymbol{p})$ varies over point light sources $\boldsymbol{p}$. Based on the description above, it is clear that the defocus PSF $k^{\boldsymbol{c}(\boldsymbol{p}),\boldsymbol{p}_d}(\boldsymbol{y},\boldsymbol{p})$ is space variant.

Thirdly, the image produced by the camera imaging system is considered. Similar to the scene intensity function $f^0(\boldsymbol{p})$, a noise-free image $g^0(\boldsymbol{y})$ in the image plane can be viewed as an intensity distribution produced by the corresponding scene intensity function $f^0(\boldsymbol{p})$. In addition, for a camera imaging system, its image plane is a 2D plane of finite physical size, so it can be described by a close set $\Gamma \in \mathbb{R}^2$. As a close subset of $\mathbb{R}^2$, Γ is measurable and its measure is positive, that is, $m(\Gamma) > 0$ [48]. Since the operator $\mathcal{A}$ is bounded, we have

$$\int_{\Gamma} \left|g^0(\boldsymbol{y})\right|^2 \mathrm{d}\boldsymbol{y} = \int_{\Gamma} \left|\mathcal{A}f^0(\boldsymbol{y})\right|^2 \mathrm{d}\boldsymbol{y} \leq K_{max} \int_{\mathbb{R}^3} \left|f^0(\boldsymbol{p})\right|^2 \mathrm{d}\boldsymbol{p} < \infty, \tag{3.5}$$

where K_{max} is an upper bound of $k(\boldsymbol{y},\boldsymbol{p})$ given in Eq. (3.4). The inequality 3.5 shows that the noise free image $g^0(\boldsymbol{y})$ is also square integrable. Therefore, the image space formed by all noise free images, denoted by $\boldsymbol{\mathcal{Y}}$, is a $\mathcal{L}^2(\Gamma)$-space and thus is a Hilbert space [48].

During the image recording process of a camera, the influence of noise should be

taken into account. For simplicity, although [27] points out that the real sensor noise is partly intensity-dependent, here the sensor noise ω is assumed to be additive and is an independent and identically distributed (i.i.d.) random variable, which follows, e.g. a Gaussian or Poison distribution. So the final captured noisy image g is given as

$$\begin{aligned} g(\boldsymbol{y}) &= g^0(\boldsymbol{y}) + \omega(\boldsymbol{y}) \\ &= \int_{\mathbb{R}^3} k^{\boldsymbol{c}(\boldsymbol{p}),\boldsymbol{p}_d}(\boldsymbol{y},\boldsymbol{p})\, f^0(\boldsymbol{p})\mathrm{d}\boldsymbol{p} + \omega(\boldsymbol{y}). \end{aligned} \tag{3.6}$$

It is worth pointing out that different from the blurring degradation, which is generally a deterministic process, the noise degradation process is stochastic, so that how a single image will be affected is undetermined [4].

For the discrete case, the image plane can be described as a 2D lattice of M pixels, then the discrete image $\boldsymbol{g_M}$ can be written as

$$\boldsymbol{g_M}[\boldsymbol{m}] = \int_{\Gamma} p_{\boldsymbol{m}}(\boldsymbol{y})\, g(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}, \tag{3.7}$$

where $\boldsymbol{m} = [m_1, m_2]^\top$ is the discrete image index, and $p_{\boldsymbol{m}}$, which represents the detector's response, is a weight kernel which is generally modelled by a rectangular function. Substituting Eq. (3.6) into Eq. (3.7), we have

$$\boldsymbol{g_M}[\boldsymbol{m}] = \int_{\Gamma}\int_{\mathbb{R}^3} p_{\boldsymbol{m}}(\boldsymbol{y})\, k^{\boldsymbol{c}(\boldsymbol{p}),\boldsymbol{p}_d}(\boldsymbol{y},\boldsymbol{p})\, f^0(\boldsymbol{p})\, \mathrm{d}\boldsymbol{p}\mathrm{d}\boldsymbol{y} + \int_{\Gamma} p_{\boldsymbol{m}}(\boldsymbol{y})\, \omega(\boldsymbol{y})\mathrm{d}\boldsymbol{y}. \tag{3.8}$$

Eq. (3.8) is a semi-discrete description of the space variant imaging system. All discrete images can be represented as vectors by, e.g. lexicographical ordering of pixels, and those image vectors form a vector space of M-dimensions, denoted by $\boldsymbol{\mathcal{Y}_M}$ [4].

Similarly, the object function f^0 can also be represented by an array of finite number of values, to make the description of a camera imaging system completely discretised. As discussed before, the 3D scene can be viewed as a point cloud. If the scene space is uniformly partitioned into N sub-spaces, and each sub-space is small enough to be represented by a single point within it, the scene is simplified to be of N point light sources. A combination of them can be thought as an approximation of the

original 3D scene as follows

$$f^0(\boldsymbol{p}) = \sum_{n=1}^{N} \boldsymbol{f_N^0}[\boldsymbol{n}]\, r_{\boldsymbol{n}}(\boldsymbol{p}), \tag{3.9}$$

where $r_{\boldsymbol{n}}$ denotes the position of the n-th point light source, e.g. $r_{\boldsymbol{n}}(\boldsymbol{p}) = \delta(\boldsymbol{p} - \boldsymbol{p_n})$, and it can be viewed as a scene where only this single point is visible. Similar to discrete images, all discrete scene intensity functions can also be represented as vectors, and all scene intensity vectors form a N-dimensional vector space, denoted by $\boldsymbol{\mathcal{X}_N}$ [4]. Substituting Eq. (3.9) into Eq. (3.7), we have a complete discrete description of the space variant imaging system, as follows

$$\begin{aligned}
\boldsymbol{g_M}[\boldsymbol{m}] &= \int_{\Gamma} p_{\boldsymbol{m}}(\boldsymbol{y}) \left(g^0(\boldsymbol{y}) + \omega(\boldsymbol{y})\right) \mathrm{d}\boldsymbol{y} \\
&= \int_{\Gamma} p_{\boldsymbol{m}}(\boldsymbol{y}) \int_{\mathbb{R}^3} k^{\boldsymbol{c}(\boldsymbol{p}),\boldsymbol{p}_d}(\boldsymbol{y},\boldsymbol{p})\, f^0(\boldsymbol{p})\, \mathrm{d}\boldsymbol{p}\mathrm{d}\boldsymbol{y} + \int_{\Gamma} p_{\boldsymbol{m}}(\boldsymbol{y})\, \omega(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y} \\
&= \int_{\Gamma} p_{\boldsymbol{m}}(\boldsymbol{y}) \int_{\mathbb{R}^3} k^{\boldsymbol{c}(\boldsymbol{p}),\boldsymbol{p}_d}(\boldsymbol{y},\boldsymbol{p}) \sum_{n=1}^{N} \boldsymbol{f_N^0}[\boldsymbol{n}]\, r_{\boldsymbol{n}}(\boldsymbol{p})\, \mathrm{d}\boldsymbol{p}\mathrm{d}\boldsymbol{y} + \int_{\Gamma} p_{\boldsymbol{m}}(\boldsymbol{y})\, \omega(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y} \\
&= \sum_{n=1}^{N} \boldsymbol{f_N^0}[\boldsymbol{n}] \int_{\Gamma} \int_{\mathbb{R}^3} p_{\boldsymbol{m}}(\boldsymbol{y})\, k^{\boldsymbol{c}(\boldsymbol{p}),\boldsymbol{p}_d}(\boldsymbol{y},\boldsymbol{p})\, r_{\boldsymbol{n}}(\boldsymbol{p})\, \mathrm{d}\boldsymbol{p}\mathrm{d}\boldsymbol{y} + \int_{\Gamma} p_{\boldsymbol{m}}(\boldsymbol{y})\, \omega(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y} \\
&= \sum_{n=1}^{N} \boldsymbol{h}^{\boldsymbol{C_N}[\boldsymbol{n}],\boldsymbol{D_N}[\boldsymbol{n}]}[\boldsymbol{m},\boldsymbol{n}]\, \boldsymbol{f_N^0}[\boldsymbol{n}] + \boldsymbol{\omega_M}[\boldsymbol{m}],
\end{aligned} \tag{3.10}$$

where $\boldsymbol{h}^{\boldsymbol{C_N}[\boldsymbol{n}],\boldsymbol{D_N}[\boldsymbol{n}]}[\boldsymbol{m},\boldsymbol{n}] = \int_{\Gamma} \int_{\mathbb{R}^3} p_{\boldsymbol{m}}(\boldsymbol{y})\, k^{\boldsymbol{c}(\boldsymbol{p}),\boldsymbol{p}_d}(\boldsymbol{y},\boldsymbol{p})\, r_{\boldsymbol{n}}(\boldsymbol{p})\, \mathrm{d}\boldsymbol{p}\mathrm{d}\boldsymbol{y}$ denotes the discrete PSF, $\boldsymbol{\omega_M}$ represents the sensor noise on the discrete image plane, and $\boldsymbol{C_N}$ and $\boldsymbol{D_N}$ are vectors representing camera settings and depths of all point light sources, respectively.

Since the process description given in Eq. (3.10) is completely discrete, it is possible to rewrite it as a matrix-vector multiplication form as suggested in [29]. As mentioned above, $\boldsymbol{g_M}$ and $\boldsymbol{\omega_M}$ are a M-dimensional noisy image vector and a noise vector, respectively, in the space $\boldsymbol{\mathcal{Y}_M}$; $\boldsymbol{f_N^0}$ is a scene intensity vector of N-dimension in the space $\boldsymbol{\mathcal{X}_N}$. Those three vectors are linked by the camera system matrix $\boldsymbol{H_{C_N,D_N}}$ of size $M \times N$, whose n-th column is the discrete PSF $\boldsymbol{h}^{\boldsymbol{C_N}[\boldsymbol{n}],\boldsymbol{D_N}[\boldsymbol{n}]}$ corresponding to the n-th point light source, with normalised unit intensity. Based on the description above, we finally have

$$\boldsymbol{g_M} = \boldsymbol{H_{C_N,D_N}} \boldsymbol{f_N^0} + \boldsymbol{\omega_M}. \tag{3.11}$$

Please notice that in most of the cases, $N \gg M$ is valid.

3.2 Space invariant imaging system

In the previous section, a camera imaging system is shown to be a space variant system in general. In this section, a special case where it can be treated as a space invariant system is derived.

As pointed out in Section 3.1, the camera imaging system is space variant, since the PSF is space variant. The reason of having a space variant PSF is two-fold. One is the complex scene structure; the other is the limited physical size of the lens. They jointly cause the problem that different point light sources have different depths and effective apertures. However, this problem does not exist in a certain situation where the scene contains merely a fronto-parallel plane. In such a situation, all points in the scene space share the same depth d and both self-occlusions and occlusions are inherently avoided, which lead to a space invariant PSF $k^{\boldsymbol{c},d}$ and thus a space invariant camera imaging system. In this case, the operator $\mathcal{A}$ can be described as a convolution, and Eq. (3.6) can be rewritten as

$$\begin{aligned} g(\boldsymbol{y}) &= g^0(\boldsymbol{y}) + \omega(\boldsymbol{y}) \\ &= \int_{\mathbb{R}^2} k^{\boldsymbol{c},d}(\boldsymbol{y}, \boldsymbol{p_x})\, f^0(\boldsymbol{p_x}) \mathrm{d}\boldsymbol{p_x} + \omega(\boldsymbol{y}) \\ &= \frac{1}{\alpha^2} \int_{\mathbb{R}^2} k^{\boldsymbol{c},d}\left(\boldsymbol{y}, \frac{\tilde{\boldsymbol{p}}_{\boldsymbol{x}}}{\alpha}\right) f^0(\frac{\tilde{\boldsymbol{p}}_{\boldsymbol{x}}}{\alpha}) \mathrm{d}\tilde{\boldsymbol{p}}_{\boldsymbol{x}} + \omega(\boldsymbol{y}), \end{aligned} \tag{3.12}$$

where $\tilde{\boldsymbol{p}}_{\boldsymbol{x}} = \alpha \boldsymbol{p_x}$ with $\alpha = \dfrac{-l_f}{d}$, representing the lens magnification, and l_f is the distance between the lens and the image plane as shown in Figure 3.1. Let $\tilde{k}^{\boldsymbol{c},d}(\boldsymbol{y}, \tilde{\boldsymbol{p}}_{\boldsymbol{x}}) \triangleq k^{\boldsymbol{c},d}\left(\boldsymbol{y}, \dfrac{\tilde{\boldsymbol{p}}_{\boldsymbol{x}}}{\alpha}\right)$ and $\tilde{f}^0(\tilde{\boldsymbol{p}}_{\boldsymbol{x}}) \triangleq f^0(\dfrac{\tilde{\boldsymbol{p}}_{\boldsymbol{x}}}{\alpha})$, Eq. (3.12) becomes

$$\begin{aligned} g(\boldsymbol{y}) &= \frac{1}{\alpha^2} \int_{\mathbb{R}^2} \tilde{k}^{\boldsymbol{c},d}(\boldsymbol{y}, \tilde{\boldsymbol{p}}_{\boldsymbol{x}})\, \tilde{f}^0(\tilde{\boldsymbol{p}}_{\boldsymbol{x}}) \mathrm{d}\tilde{\boldsymbol{p}}_{\boldsymbol{x}} + \omega(\boldsymbol{y}) \\ &= \frac{1}{\alpha^2} \int_{\mathbb{R}^2} \tilde{k}^{\boldsymbol{c},d}(\boldsymbol{y} - \tilde{\boldsymbol{p}}_{\boldsymbol{x}})\, \tilde{f}^0(\tilde{\boldsymbol{p}}_{\boldsymbol{x}}) \mathrm{d}\tilde{\boldsymbol{p}}_{\boldsymbol{x}} + \omega(\boldsymbol{y}). \end{aligned} \tag{3.13}$$

Thus, Eq. (3.13) can be simply given as

$$g = \frac{1}{\alpha^2} \tilde{k}^{\boldsymbol{c},d} \otimes \tilde{f}^0 + \omega, \tag{3.14}$$

where $\otimes$ denotes convolution.

The same analysis done in the case of the space variant system in Section 3.1 can be applied here to describe a completely discrete space invariant system, as follows

$$\boldsymbol{g_M} = \frac{1}{\alpha^2}\tilde{\boldsymbol{H}}_{\boldsymbol{c},d}\tilde{\boldsymbol{f}}^0_{\boldsymbol{N}} + \boldsymbol{\omega_M}. \tag{3.15}$$

Notice that now the system matrix $\tilde{\boldsymbol{H}}$ is characterised by a single discrete PSF $\tilde{\boldsymbol{h}}^{\boldsymbol{c},d}$.

In real cases, the aforementioned situation is in fact rare. However, although it is often unrealistic to treat the whole camera imaging system as space invariant, locally it can be valid if a mild assumption is made. This assumption is that in most of the cases, the structure of a 3D scene can be treated as piece-wise planar. More specifically, it means that the PSF within a small sub-domain $\mathcal{D}_\Gamma$ of the image plane Γ is space invariant, if its corresponding limited sub-domain $\mathcal{D}_{\mathbb{R}^2}$ in the scene plane $\mathbb{R}^2$ can be treated as a fronto-parallel plane [4]. Therefore, if we partition the image plane Γ into multiple small sub-domains $\{\mathcal{D}_{\Gamma_l}\}$ where the PSF is space-invariant. For each $\mathcal{D}_{\Gamma_l}$, we have

$$g(\boldsymbol{y}) = \frac{1}{\alpha^2}\int\limits_{\mathcal{D}_{\mathbb{R}^2_l}} \tilde{k}^{\boldsymbol{c}_l,d_l}\left(\boldsymbol{y} - \tilde{\boldsymbol{p}}_{\boldsymbol{x}}\right)\tilde{f}^0(\tilde{\boldsymbol{p}}_{\boldsymbol{x}})\mathrm{d}\tilde{\boldsymbol{p}}_{\boldsymbol{x}} + \omega(\boldsymbol{y}), \forall \boldsymbol{y} \in \mathcal{D}_{\Gamma_l}, \tag{3.16}$$

where $\mathcal{D}_{\mathbb{R}^2_l}$ is the corresponding sub-domain of $\mathcal{D}_{\Gamma_l}$ in the scene plane $\mathbb{R}^2$. Similarly, locally the completely discrete description is given as follows,

$$\boldsymbol{g}_{\boldsymbol{L_M}} = \frac{1}{\alpha^2}\tilde{\boldsymbol{h}}^{\boldsymbol{c}_l,d_l} \otimes \tilde{\boldsymbol{f}}^0_{\boldsymbol{L_N}} + \boldsymbol{\omega}_{\boldsymbol{L_M}}, \tag{3.17}$$

where $\boldsymbol{L_M}$ and $\boldsymbol{L_N}$ represent corresponding sub-domains in the discrete image plane and the discrete scene plane, respectively.

For the rest part of the thesis, the scene intensity function f^0 is assumed to be already scaled such that $\alpha = 1$, and thus the notation $\sim$ can be ignored for simplicity.

3.3 Aperture superposition principle

In this section, the camera imaging system is presented from another point of view based on the aperture superposition principle.

Despite the accuracy, it might be computationally hard to apply the previous description of a camera imaging system for complex scenes. Lanman et al. [20] showed

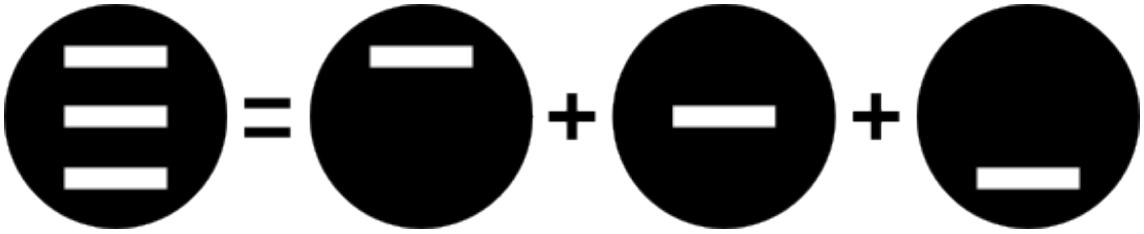


Figure 3.3 *An example of aperture superposition.*

that an aperture can be equivalently viewed as a superposition of a set of elementary apertures, and one example is shown in Figure 3.3. The image captured with the whole aperture can be approximated by a superposition of a set of images captured with those elementary apertures, which is named as aperture superposition principle. Mathematically, it can be expressed as

$$\boldsymbol{g}_{\boldsymbol{M}} = \sum_i a_i \boldsymbol{g}_{\boldsymbol{M}_i}, \tag{3.18}$$

where a_i is the transmission efficiency of the i-th elementary aperture and $\boldsymbol{g}_{\boldsymbol{M}_i}$ is the image captured with this elementary aperture.

Ideally, any aperture pattern can be divided into a set of 'pinholes', and each image captured with a 'pinhole' aperture is all-in-focus. By doing so, calculating PSFs is totally avoided. Also, the occlusion problem is automatically solved since a point light source does not appear in the all-in-focus image for a particular view if it is blocked. But in practice, beside those advantages, this method also has its own drawbacks. A real pinhole aperture will cause significant diffraction effects that should not be ignored, so in order to keep the diffraction effects negligible, a 'pinhole' aperture must be big enough. However, if a 'pinhole' aperture has too big opening, it will not lead to a all-in-focus image due to the lens effects. Thus, care must be taken when choosing the size of a 'pinhole', to keep a balance between minimising diffraction effects and minimising lens defocus blur effects.

Overall, if a good pattern partitioning resolution is selected, this method works sufficiently well and can be conveniently used in many applications.

4. DEPTH FROM DEFOCUS

In Chapter 3, a camera imaging system is analysed, which shows how an image of a 3D scene is formed. This image formation process is known as a direct process from cause to effect, or from a rich information state to a poor information state. In this chapter, however, its inverse problem, the problem of estimating scene information based on a limited number of captured images, is targeted. More specifically, the depth information is of particular interest, and the defocus blur cue is chosen to be the main depth cue of interest, and thus the problem of this type is specified as finding depth from defocus (DfD). This inverse problem is quite challenging, due to the information lost during the direct process. This chapter begins with defining and analysing the problem. For solving the mentioned problem, existing methods based on two solving strategies are introduced and discussed.

4.1 Problem statement and analysis

The problem of DfD can be expressed as: given N images $\{\boldsymbol{g}_{\boldsymbol{M}_n} | 1 \leq n \leq N, n \in \mathbb{Z}^+, N \in \mathbb{Z}^+\}$ captured with known camera settings from the same view, how to extract the defocus information encoded in images and use it to do depth estimation. Particularly, within this thesis, changing camera setting is restricted to changing the aperture shape. In addition, N is limited to either 1 or 2, for practical usage considerations.

Since it is assumed that both the camera and 3D scene are fixed, the depth information remains unchanged in all images. As shown in the image formation model given in 3.11, the depth information is encoded according to PSFs. In addition, the model also shows that the depth information $\boldsymbol{D_N}$ is independent to the scene intensity function $\boldsymbol{f_N^0}$, which is also unknown, and estimating $\boldsymbol{f_N^0}$ is known as the problem of the image restoration. Nevertheless, both problems are challenging since they are ill-posed in the sense of Hadamard, who suggested that a physically meaningful model should satisfy three properties [4]:
1) a solution exists;
2) the solution is unique;

3) the solution depends continuously on initial conditions.
This ill-posedness comes due to the fact that the scene information is not completely recorded in images, which is best viewed in the continuous case. Mainly there are two reasons of losing information. One is that a camera imaging system is band-limited. It means that in the frequency domain, the optical transfer function (OTF) of a camera, denoted by $K(\boldsymbol{\xi})$, tends to zeros in the high frequency zone, due to the finite size of imaging lens. The other is that even within the band of $K(\boldsymbol{\xi})$, it may have zeros at certain frequencies. Consequently, g^0, as a degraded representation of f^0, does not contain complete information any more, since in the frequency domain, we have

$$G^0(\boldsymbol{\xi}) = K(\boldsymbol{\xi}) F^0(\boldsymbol{\xi}). \tag{4.1}$$

As a consequence of this incompleteness, there are multiple pairs of K and F^0 that satisfy Eq. (4.1). For example, when $G^0(\boldsymbol{\xi}) = 0$ for certain $\boldsymbol{\xi}$, it may be $K(\boldsymbol{\xi}) = 0$ or $F^0(\boldsymbol{\xi}) = 0$, or both. Clearly, this makes both depth estimation and image restoration impossible to be solved uniquely, so it violates the second condition of Hadamard, which makes the problems ill-posed [4].

How to treat $\boldsymbol{f_N^0}$ lead to two categories of DfD solving strategies. One solves depth estimation and image restoration simultaneously since both are demanded in many applications; while the other bypasses the image restoration and directly focuses on the depth estimation.

Regarding the resolution, since both $\boldsymbol{D_N}$ and $\boldsymbol{f_N^0}$ are recorded with the same image resolution, during the discussion within this thesis, both problems are solved on the image grid including all pixels. The former, an estimated depth information of image resolution, is known as the dense depth map and is denoted by $\boldsymbol{D_M}$; the latter, a restored scene intensity function of image resolution, is viewed as the all-in-focus image and is denoted by $\boldsymbol{f_M}$.

4.2 Solving strategies: restoration based

In this section, a class of methods that follow the restoration-based strategy are introduced. In general, methods following this strategy try to obtain the depth map and the restored image simultaneously, and very often the quality of estimated depth map depends on the quality of the restored image and vice versa.

The problems are usually analysed by using Bayesian methods, for two reasons. One is that the formation of an image is a random process due to the existence of noise, so

it is natural to use statistical methods to treat the problem. The other is that since both problems are ill-posed due to the incompleteness of information, additional information, or constraints, must be introduced as a compensation, and Bayesian methods are convenient for allowing introducing complex *a priori* information, e.g. information that is hard to be explicitly given in formulae.

Under the Bayesian method, the depth map $\boldsymbol{D_M}$ and the all-in-focus image $\boldsymbol{f_M}$ as well as the captured image $\boldsymbol{g_M}$ and noise $\boldsymbol{\omega_M}$ are all viewed as random variables with probability distributions, denoted by $p(\boldsymbol{D_M})$, $p(\boldsymbol{f_M})$, $p(\boldsymbol{g_M})$ and $p_{\boldsymbol{\omega_M}}(\boldsymbol{\omega_M})$, respectively. Particularly, the joint distribution of $\boldsymbol{D_M}$, $\boldsymbol{f_M}$ and $\boldsymbol{g_M}$, denoted by $p(\boldsymbol{D_M}, \boldsymbol{f_M}, \boldsymbol{g_M})$, gives a complete probabilistic description of the whole system, since it covers all variables of interest. By using Bayes' rule, we have

$$\begin{aligned} p(\boldsymbol{D_M}, \boldsymbol{f_M}, \boldsymbol{g_M}) &= p(\boldsymbol{D_M}, \boldsymbol{f_M}|\boldsymbol{g_M})\, p(\boldsymbol{g_M}) \\ &= p(\boldsymbol{g_M}|\boldsymbol{D_M}, \boldsymbol{f_M})\, p(\boldsymbol{D_M})\, p(\boldsymbol{f_M}). \end{aligned} \tag{4.2}$$

When captured images, which are observations of variable $\boldsymbol{g_M}$, are taken into account, we have

$$p\left(\boldsymbol{D}, \boldsymbol{f_M}|\boldsymbol{g}_{\boldsymbol{M}_{1,\ldots,N}}\right) \propto p\left(\boldsymbol{g}_{\boldsymbol{M}_{1,\ldots,N}}|\boldsymbol{D_M}, \boldsymbol{f_M}\right) p(\boldsymbol{D_M})\, p(\boldsymbol{f_M}). \tag{4.3}$$

In Eq. (4.3), $p(\boldsymbol{D_M})$ and $p(\boldsymbol{f_M})$ are prior distributions of $\boldsymbol{D_M}$ and $\boldsymbol{f_M}$, respectively. They contain *a priori* information and thus introduce additional constraints to the system. $p\left(\boldsymbol{g}_{\boldsymbol{M}_{1,\ldots,N}}|\boldsymbol{D_M}, \boldsymbol{f_M}\right)$ is the likelihood measuring the probability that images are generated by the scene information $\boldsymbol{D_M}$ and $\boldsymbol{f_M}$. Finally, $p\left(\boldsymbol{D}, \boldsymbol{f_M}|\boldsymbol{g}_{\boldsymbol{M}_{1,\ldots,N}}\right)$ is known as the joint posterior distribution of $\boldsymbol{D_M}$ and $\boldsymbol{f_M}$, and it is the distribution of interest since the pair $\{\boldsymbol{D_M}^*, \boldsymbol{f_M}^*\}$ maximising this distribution is considered as the best solution of the problem. That is, the problem is presented as a maximum *a posteriori* (MAP) probability estimation,

$$\begin{aligned} \boldsymbol{D_M}^*, \boldsymbol{f_M}^* &= \underset{\boldsymbol{D_M}, \boldsymbol{f_M}}{\arg\max}\, p\left(\boldsymbol{D_M}, \boldsymbol{f_M}|\boldsymbol{g}_{\boldsymbol{M}_{1,\ldots,N}}\right) \\ &= \underset{\boldsymbol{D_M}, \boldsymbol{f_M}}{\arg\max}\, p\left(\boldsymbol{g}_{\boldsymbol{M}_{1,\ldots,N}}|\boldsymbol{D_M}, \boldsymbol{f_M}\right) p(\boldsymbol{D_M})\, p(\boldsymbol{f_M}) \\ &= \underset{\boldsymbol{D_M}, \boldsymbol{f_M}}{\arg\max} \prod_{n=1}^{N} \left\{p\left(\boldsymbol{g}_{\boldsymbol{M}_n}|\boldsymbol{D_M}, \boldsymbol{f_M}\right)\right\} p(\boldsymbol{D_M})\, p(\boldsymbol{f_M}) \\ &= \underset{\boldsymbol{D_M}, \boldsymbol{f_M}}{\arg\max} \prod_{n=1}^{N} \left\{p_{\boldsymbol{\omega_M}}\left(\boldsymbol{g}_{\boldsymbol{M}_n} - \boldsymbol{H}_{\boldsymbol{C}_{\boldsymbol{M}_n}, \boldsymbol{D_M}} \boldsymbol{f_M}\right)\right\} p(\boldsymbol{D_M})\, p(\boldsymbol{f_M}), \end{aligned} \tag{4.4}$$

where Eq. (4.4) is obtained by using the model given in Eq. (3.11), which implies

that

$$p\left(\boldsymbol{g}_{\boldsymbol{M}_n}|\boldsymbol{D}_{\boldsymbol{M}},\boldsymbol{f}_{\boldsymbol{M}}\right) = p_{\boldsymbol{\omega}_{\boldsymbol{M}}}\left(\boldsymbol{\omega}_{\boldsymbol{M}}\right) = p_{\boldsymbol{\omega}_{\boldsymbol{M}}}\left(\boldsymbol{g}_{\boldsymbol{M}_n} - \boldsymbol{H}_{\boldsymbol{C}_{\boldsymbol{M}_n},\boldsymbol{D}_{\boldsymbol{M}}}\boldsymbol{f}_{\boldsymbol{M}}\right), \tag{4.5}$$

and $\boldsymbol{C_M}$ denotes effective camera settings and is defined in a similar way to $\boldsymbol{D_M}$.

The function in Eq. (4.4) can be maximised directly if proper distributions are chosen, and it gives a global solution for the problem, c.f. [38]. However, directly acquiring global solutions requires an explicit mathematical model of PSF, which may not be accurately known in certain cases. A more accurate way is to work on PSFs captured at a finite set of pre-sampled depths $\mathcal{K}$, since experimentally modelled PSFs are of better accuracy [10]. Moreover, taking the advantage of locally space invariant assumption made in Section 3.2, in a sub-domain $\boldsymbol{L}$, i.e. a square patch centred in the l-th pixel, the system can be treated as space invariant and thus $\boldsymbol{C_L}$ and $\boldsymbol{D_L}$ are determined to be uniform. Since no occlusion exists and the camera setting $\boldsymbol{C_L}$ is assumed to be known (see Section 4.1), $\boldsymbol{D_L}$ and $\boldsymbol{H}_{\boldsymbol{C_L},\boldsymbol{D_L}}$ form am one-to-one mapping. That is, locally estimating depth is equivalent to determining the correct PSF, which simplifies the problem to a large extent. Therefore, the problem stated in Eq. (4.4) is to be solved patch-wisely, as follows,

$$\boldsymbol{D_M}^*[l], \boldsymbol{f_M}^*[l] = \underset{d_k,\boldsymbol{f_M}}{\arg\max} \sum_{\boldsymbol{L}} \prod_{n=1}^{N} \left\{p_{\boldsymbol{\omega}_{\boldsymbol{M}}}\left(\boldsymbol{g}_{\boldsymbol{M}_n} - \boldsymbol{h}^{\boldsymbol{c}_n,d_k} \otimes \boldsymbol{f_M}\right)\right\} p\left(\boldsymbol{f_M}\right), d_k \in \mathcal{K}. \tag{4.6}$$

Notice that $p\left(\boldsymbol{D_M}\right)$ is dropped since within the patch L, it is a constant.

In order to solve Eq. (4.6), proper probability distributions must be chosen. In most of the cases, the noise $\boldsymbol{\omega_M}$ can be assumed to be a multivariate white Gaussian noise with distribution $\boldsymbol{\omega_M} \sim \mathcal{N}\left(0, \sigma^2\boldsymbol{I}\right)$, where σ^2 is the noise variance and $\boldsymbol{I}$ represents an identity matrix. Therefore, we have $\boldsymbol{g_M} \sim \mathcal{N}\left(\boldsymbol{H}_{\boldsymbol{C_M},\boldsymbol{D_M}}\boldsymbol{f_M}, \sigma^2\boldsymbol{I}\right)$. However, for choosing the prior distribution $p\left(\boldsymbol{f_M}\right)$, care must be taken. A good prior should reflect the properties that a potential solution should have. For the image prior selection, one good way is to use natural image statistics. Statistics show that in the spatial domain, the output obtained by applying derivative-like filters on a natural image form a distribution that is peaked at zero and heavy tailed, which means that natural images are more likely to be smooth and have sparse edges. In the frequency domain, the power spectra of natural images tend to be dominated by the low frequency components and the weights of frequency components fall off as $\frac{1}{\boldsymbol{\xi}^2}$, and this is known as the $\frac{1}{\boldsymbol{\xi}}$ law [57]. Those two statistical observations are

consistent, since sharper edges correspond to higher frequency components.

Two examples of image prior consistent with statistical observations are given by Levin et al. [23] and Zhou et al. [60] in the spatial domain and the frequency domain, respectively. In [23], a sparse derivatives prior is designed as

$$p\left(\boldsymbol{f_M}\right) \propto \exp\left(-\frac{\rho\left(\nabla_v \boldsymbol{f_M}\right)+\rho\left(\nabla_h \boldsymbol{f_M}\right)}{2}\right), \tag{4.7}$$

where ∇_v and ∇_h are derivative operators taking the gradient of image in the vertical direction and the horizontal direction, respectively; and ρ is selected to be a function with heavy-tail, for example, $\rho\left(\boldsymbol{z}\right)=\|\boldsymbol{z}\|_{0.8}^{0.8}$, where $\|\ \|_p$ denotes the p-norm. On the other hand, in [60], an image prior given in the frequency domain is directly learnt from a set of natural images. The image prior used in [60] is of the type

$$p\left(\boldsymbol{F_M}\right) \propto \exp\left(-0.5\left\|\boldsymbol{\Psi}\bullet\boldsymbol{F_M}\right\|_2^2\right), \tag{4.8}$$

where $\bullet$ denotes the element-wise multiplication, $\boldsymbol{\Psi}$ is a linear weight matrix, and it can be learnt as

$$\left|\boldsymbol{\Psi}\left(\boldsymbol{\xi}\right)\right|^2=\frac{1}{\int\limits_{\boldsymbol{F_M^0}}\left|\boldsymbol{F_M^0}\left(\boldsymbol{\xi}\right)\right|^2\mu\left(\boldsymbol{F_M^0}\right)}, \tag{4.9}$$

where $|\ |^2$ denotes the element-wise square operation and $\mu\left(\boldsymbol{F_M^0}\right)$ is the possibility measure of the discrete Fourier transform (DFT) of an all-in-focus image $\boldsymbol{f_M^0}$. In the spatial domain, by applying $\boldsymbol{g_M}\sim\mathcal{N}\left(\boldsymbol{H_{C_M,D_M}f_M},\sigma^2\boldsymbol{I}\right)$ and Levin's image prior given in Eq. (4.7) to the problem described in Eq. (4.6), we have

$$\begin{aligned}
\boldsymbol{D_M}^*[l],\boldsymbol{f_M}^*[l] &= \underset{d_k,\boldsymbol{f_M}}{\arg\max}\sum_{\boldsymbol{L}}\prod_{n=1}^{N}\left\{p_{\boldsymbol{\omega_M}}\left(\boldsymbol{g}_{\boldsymbol{M}_n}-\boldsymbol{H}_{\boldsymbol{c}_n,d_k}\boldsymbol{f_M}\right)\right\}p\left(\boldsymbol{f_M}\right)\\
&= \underset{d_k,\boldsymbol{f_M}}{\arg\max}\sum_{\boldsymbol{L}}\prod_{n=1}^{N}\left\{e^{\left(\frac{-0.5}{\sigma^2}\left\|\boldsymbol{g}_{\boldsymbol{M}_n}-\boldsymbol{H}_{\boldsymbol{c}_n,d_k}\boldsymbol{f_M}\right\|_2^2\right)}\right\}e^{\left(-\frac{\rho\left(\nabla_v\boldsymbol{f_M}\right)+\rho\left(\nabla_h\boldsymbol{f_M}\right)}{2}\right)}\\
&= \underset{d_k,\boldsymbol{f_M}}{\arg\min}\sum_{\boldsymbol{L}}\sum_{n=1}^{N}\left(\left\|\boldsymbol{g}_{\boldsymbol{M}_n}-\boldsymbol{H}_{\boldsymbol{c}_n,d_k}\boldsymbol{f_M}\right\|_2^2\right)+\sigma^2\left(\rho\left(\nabla_v\boldsymbol{f_M}\right)+\rho\left(\nabla_h\boldsymbol{f_M}\right)\right).
\end{aligned} \tag{4.10}$$

Please notice that the selection of both noise and image prior are not necessarily restricted to be from the exponential family. However, such choices make the analytical derivation possible, as shown in Eq. (4.10). Here it is worth mentioning that

the inclusion of the prior information in Bayesian methods serves as a regularisation.

Although the problem given in Eq. (4.10) is clear, its solution is hard to acquire. A general procedure is to separate the problem into two parts [23], [53]. In the first part, the aim is to acquire a restored image $\hat{\boldsymbol{f}}_{\boldsymbol{M}}^{k}$ for each given depth $d_k \in \mathcal{K}$, as

$$\hat{\boldsymbol{f}}_{\boldsymbol{M}}^{k} = \arg\min_{\boldsymbol{f}_{\boldsymbol{M}}} \sum_{n=1}^{N} \left(\left\| \boldsymbol{g}_{\boldsymbol{M}_n} - \boldsymbol{H}_{\boldsymbol{c}_n,d_k} \boldsymbol{f}_{\boldsymbol{M}} \right\|_2^2 \right) + \sigma^2 \left(\rho \left(\nabla_v \boldsymbol{f}_{\boldsymbol{M}} \right) + \rho \left(\nabla_h \boldsymbol{f}_{\boldsymbol{M}} \right) \right), \quad (4.11)$$

and it can be minimised by using, e.g. iterative re-weighted least squares(IRLS) algorithms [22] in the spatial domain.

In the second part, those restored images $\hat{\boldsymbol{f}}_{\boldsymbol{M}}^{k}$ and corresponding pre-sampled PSF sets $\left\{ \boldsymbol{h}^{\boldsymbol{c}_n,d_k} \right\}$ are put in pairs. For example, the k-th pair is $\left\{ \hat{\boldsymbol{f}}_{\boldsymbol{M}}^{k}, \left\{ \boldsymbol{h}^{\boldsymbol{c}_n,d_k} \right\} \right\}$. The estimated local depth map and the local restored image are selected from those pairs, and the pair that locally maximises the likelihood function is thought to be the optimal choice. This maximum likelihood estimation (MLE) is shown as follows,

$$\begin{aligned} \boldsymbol{D_M}^*[l], \boldsymbol{f_M}^*[l] &= \arg\max_{d_k, \hat{\boldsymbol{f}}_{\boldsymbol{M}}^{k}} \sum_{\boldsymbol{L}} p \left(\boldsymbol{g}_{\boldsymbol{M}_{1,\ldots,N}} | d_k, \hat{\boldsymbol{f}}_{\boldsymbol{M}}^{k} \right) \\ &= \arg\max_{d_k, \hat{\boldsymbol{f}}_{\boldsymbol{M}}^{k}} \sum_{\boldsymbol{L}} p \left(\boldsymbol{g}_{\boldsymbol{M}_{1,\ldots,N}} | \boldsymbol{h}^{\boldsymbol{c}_1,d_k}, \ldots, \boldsymbol{h}^{\boldsymbol{c}_N,d_k}, \hat{\boldsymbol{f}}_{\boldsymbol{M}}^{k} \right) \\ &= \arg\min_{d_k, \hat{\boldsymbol{f}}_{\boldsymbol{M}}^{k}} \sum_{\boldsymbol{L}} \sum_{n=1}^{N} \left(\left\| \boldsymbol{g}_{\boldsymbol{M}_n} - \boldsymbol{H}_{\boldsymbol{c}_n,d_k} \hat{\boldsymbol{f}}_{\boldsymbol{M}}^{k} \right\|_2^2 \right). \end{aligned} \quad (4.12)$$

This procedure is repeated for all pixels.

The problem described in Eq. (4.6) can also be solved in the frequency domain, as presented in [60]. According to the Parseval's theorem, the Fourier transform is an unitary operator. For the discrete case, the following relation exists,

$$\left\| \boldsymbol{z}\left[\boldsymbol{n}\right] \right\|_2^2 = \frac{1}{N} \left\| \boldsymbol{Z}\left(\boldsymbol{\xi}\right) \right\|_2^2. \quad (4.13)$$

By applying Eq. (4.13) to Eq. (4.10) and replacing image prior with Eq. (4.8), in the frequency domain, the problem becomes

$$\boldsymbol{D_M}^*[l], \boldsymbol{f_M}^*[l] = \arg\min_{d_k, \boldsymbol{f}_{\boldsymbol{M}}} \sum_{\boldsymbol{L}} \sum_{n=1}^{N} \left(\left\| \boldsymbol{G}_{\boldsymbol{M}_n} - \boldsymbol{H}^{\boldsymbol{c}_n,d_k} \bullet \boldsymbol{F}_{\boldsymbol{M}} \right\|_2^2 \right) + \left\| \boldsymbol{W} \bullet \boldsymbol{F}_{\boldsymbol{M}} \right\|_2^2, d_k \in \mathcal{K}, \quad (4.14)$$

where $\boldsymbol{W} = \sigma \boldsymbol{\Psi}$.

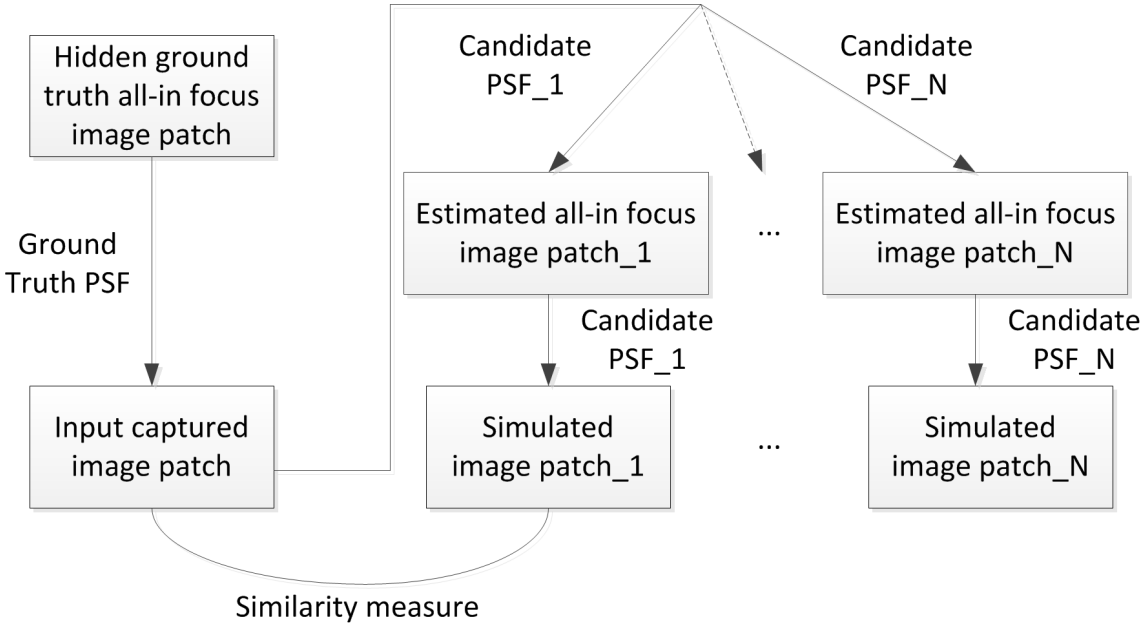


Figure 4.1 *Illustration of the principle of restoration-based strategy.*

To solve 4.14, the same procedure is used. In the first part, the aim is to acquire the DFT of the restored image $\hat{\boldsymbol{F}}_{\boldsymbol{M}}^{k}$ for each depth $d_k \in \mathcal{K}$, as

$$\hat{\boldsymbol{F}}_{\boldsymbol{M}}^{k} = \underset{\boldsymbol{F_M}}{\arg\min} \sum_{n=1}^{N} \left(\left\| \boldsymbol{G}_{\boldsymbol{M}_n} - \boldsymbol{H}^{\boldsymbol{c}_n,d_k} \bullet \boldsymbol{F}_{\boldsymbol{M}} \right\|_2^2 \right) + \left\| \boldsymbol{W} \bullet \boldsymbol{F}_{\boldsymbol{M}} \right\|_2^2, \tag{4.15}$$

and it can be solved by using a (generalised) Wiener filter in the frequency domain [60], as

$$\hat{\boldsymbol{F}}_{\boldsymbol{M}}^{k} = \frac{\sum_n \boldsymbol{G}_{\boldsymbol{M}_n} \bullet \bar{\boldsymbol{H}}^{\boldsymbol{c}_n,d_k}}{\sum_n |\boldsymbol{H}^{\boldsymbol{c}_n,d_k}|^2 + |\boldsymbol{W}|^2}, \tag{4.16}$$

where $|\boldsymbol{W}|^2$ is a matrix representing the noise-to-signal ratio (NSR).

In the second part, those DFTs $\hat{\boldsymbol{F}}_{\boldsymbol{M}}^{k}$ of restored images are transferred back to the spatial domain as $\hat{\boldsymbol{f}}_{\boldsymbol{M}}^{k}$ and associated with the corresponding pre-sampled PSF set $\left\{\boldsymbol{h}^{\boldsymbol{c}_n,d_k}\right\}$ at depth d_k. Then a MLE is solved on those pairs $\left\{\hat{\boldsymbol{f}}_{\boldsymbol{M}}^{k}, \left\{\boldsymbol{h}^{\boldsymbol{c}_n,d_k}\right\}\right\}$ as follows,

$$\boldsymbol{D_M}^*[l], \boldsymbol{f_M}^*[l] = \underset{d_k, \hat{\boldsymbol{f}}_{\boldsymbol{M}}^{k}}{\arg\min} \sum_{\boldsymbol{L}} \sum_{n=1}^{N} \left(\left\| \boldsymbol{g}_{\boldsymbol{M}_n} - \boldsymbol{h}^{\boldsymbol{c}_n,d_k} \otimes \hat{\boldsymbol{f}}_{\boldsymbol{M}}^{k} \right\|_2^2 \right). \tag{4.17}$$

This procedure is repeated for all pixels.

The procedure of algorithms following the restoration-based solving strategy is clear. As shown in Figure 4.1, a captured image patch can be thought as generated by the ground truth all-in-focus image patch and PSF. Ideally, when the correct PSF is selected, the estimated image patch and this PSF should be able to produce a simulated image patch that is similar to the captured one, measured according to a certain criterion. If an incorrect PSF is used, errors will be introduced in both the image restoration step and the image simulation step, and thus the simulated image patch will be less similar to the captured one. So it is obvious that the quality of the restored image is important. Thus, a failure in the image restoration produces erroneous results.

4.3 Solving strategies: restoration free

In Section 4.2, depth estimation is achieved based on the result of image restoration. However, it is not strictly necessary to restore the all-in-focus image, since the defocus blur cue is encoded by the PSF, which is independent to the all-in-focus image. Thus, in this section, a restoration-free strategy is applied to directly solve the problem of depth map estimation, and two ideas under this strategy are introduced. Please notice that during the discussion below, only a single image is used for notation simplicity.

The principle behind this restoration-free strategy is that different PSFs have different power spectra, which modify scene intensity functions differently, e.g. eliminating different frequency components. Consequently, corresponding noise-free images of scene intensity functions modified by the same PSF share a common frequency support, which is defined as all frequencies with non-zero responses.

Two ideas emerge about utilising the frequency supports of PSFs. One is that noise-free images degraded by the same PSF share a common frequency support and thus form a subspace of the image space. If the noise does not exist, depth estimation can be done by finding the most suitable subspace where each image patch lies. However, in most of the cases, it is unrealistic to ignore noise. Noise, which is not band-limited, can randomly change the power spectrum of an image and thus make the image deviated from its subspace. Especially, the influence of noise is heavy on frequency components where they should be zeros or negligible values, which are actually important and utilised in depth estimation, as will be discussed in Section 5.1. Therefore, in order to apply this idea, a band-limiting operator $\mathcal{P}_{\mathcal{B}}$ projecting the image patch to the frequency support $\mathcal{B}$ of a PSF is needed [4].

For a PSF $\boldsymbol{h}^{\boldsymbol{c},d}$ at depth d, the corresponding band-limiting operator $\mathcal{P}_{\mathcal{B}}^{\boldsymbol{c},d}$ can be

implemented as a filter designed either in the frequency domain or in the spatial domain. In the frequency domain, a filter $\boldsymbol{P}^{\boldsymbol{c},d}$ is proposed by Lin et al. [26], as

$$\boldsymbol{P}^{\boldsymbol{c},d}(\boldsymbol{\xi}) = \begin{cases} 1, & \boldsymbol{H}^{\boldsymbol{c},d}(\boldsymbol{\xi}) \neq 0 \\ 0, & \boldsymbol{H}^{\boldsymbol{c},d}(\boldsymbol{\xi}) = 0, \end{cases} \tag{4.18}$$

where $\boldsymbol{H}^{\boldsymbol{c},d}(\boldsymbol{\xi})$ is the DFT of the PSF $\boldsymbol{h}^{\boldsymbol{c},d}$.

In the spatial domain, on the other hand, instead of finding an operator projecting image patches to the subspace, Martinello and Favaro [29] attempt to find an operator projecting image patches to the orthogonal space of the subspace defined by a PSF. For a PSF $\boldsymbol{h}^{\boldsymbol{c},d}$ at depth d, the corresponding orthogonal operator is denoted by $\boldsymbol{H}^{\perp}_{\boldsymbol{c},d}$ and it can be learnt from training images. Given a set of all-in-focus images of the same size, when they are blurred by the same PSF, the resulting set of noise-free images will be all in the subspace defined by this PSF. If we arrange all those all-in-focus images for training in a matrix $\boldsymbol{F}^0_{train}$, where each column of it is an image vector, the noise-free defocused images can also be represented as a matrix $\boldsymbol{G}^0_{train,d}$, and

$$\boldsymbol{G}^0_{train,d} = \boldsymbol{H}_{\boldsymbol{c},d}\boldsymbol{F}^0_{train}. \tag{4.19}$$

Particularly, when the training set is sufficiently large, it can be assumed that columns of $\boldsymbol{G}^0_{train,d}$ span the subspace defined by the PSF $\boldsymbol{h}^{\boldsymbol{c},d}$ [12]. Since $\boldsymbol{H}^{\perp}_{\boldsymbol{c},d}$ projects image vectors onto the subspace perpendicular to the subspace defined by $\boldsymbol{H}_{\boldsymbol{c},d}$, we should have $\left\|\boldsymbol{H}^{\perp}_{\boldsymbol{c},d}\boldsymbol{G}^0\right\|^2 = 0$. By applying singular value decomposition (SVD) on $\boldsymbol{G}^0_{train,d}$, we have $\boldsymbol{G}^0_{train,d} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^*$, where $\boldsymbol{S}$ contains singular values, and they are assumed to be sorted as from the largest value to the smallest value. Then the matrix $\boldsymbol{U}$ can be separated into two parts like $\boldsymbol{U} = [\boldsymbol{U}_+, \boldsymbol{U}_0]$ in accordance with the corresponding singular values, where $\boldsymbol{U}_0$ corresponds to close to zero, or negligible, singular values. Therefore, $\boldsymbol{H}^{\perp}_{\boldsymbol{c},d}$ can be defined as

$$\boldsymbol{H}^{\perp}_{\boldsymbol{c},d} = \boldsymbol{U}_0\boldsymbol{U}_0^T. \tag{4.20}$$

It is important to notice that, since the resulting $\boldsymbol{H}^{\perp}_{\boldsymbol{c},d}$ is learnt from training images, when the size of training images is sufficiently large, it inherently contains image statistics results [29] that serve as a regularisation as discussed in Section 4.2.

Similar to the procedure presented in Section 4.2, depth estimation is solved pixel-wise with PSFs pre-sampled at a finite set of depths $\mathcal{K}$, and it is also done in two parts.

The first part is to construct filters at all depth $d_k \in \mathcal{K}$. For each PSF $\boldsymbol{h}^{\boldsymbol{c},d_k}$, the corresponding filter can be constructed in the frequency domain denoted by $\boldsymbol{P}^{\boldsymbol{c},d_k}(\boldsymbol{\xi})$ as shown in Eq. (4.18), or in the spatial domain denoted by $\boldsymbol{H}^{\perp}_{\boldsymbol{c},d_k}$ using Eq. (4.20).

Then in the second part, constructed filters are applied to each image patch $\boldsymbol{g_L}$, which is centred in the l-th pixel, of the same size as training images, and the one leading to the minimum residual error indicates the most suitable subspace of this image patch. That is,

$$\boldsymbol{D_M}^*[l] = \arg\min_{d_k} \sum_{\boldsymbol{L}} \left\| \boldsymbol{g_M} - \boldsymbol{p}^{\boldsymbol{c},d_k} \otimes \boldsymbol{g_M} \right\|_2^2, \tag{4.21}$$

where $\boldsymbol{p}^{\boldsymbol{c},d_k}$ is inverse DFT of $\boldsymbol{P}^{\boldsymbol{c},d}(\boldsymbol{\xi})$; or,

$$\boldsymbol{D_M}^*[l] = \arg\min_{d_k} \left\| \boldsymbol{H}^{\perp}_{d_k} \boldsymbol{g_L} \right\|_2^2. \tag{4.22}$$

This procedure is repeated for all pixels.

In the first idea, depth estimation is done by finding the most suitable subspace for an image. However, instead of utilising the whole subspace, a few of features may be enough to distinguish images modified by different PSFs. This is the second idea under the restoration-free strategy, and it can be done by using local frequency component analysis, as suggested by e.g. [7], [62], [6].

In this case, under the locally space invariant assumption, the depth estimation is formulated as a MLE problem as follows

$$\boldsymbol{D}^* = \arg\max_{d} p\left(\boldsymbol{\mathcal{R}} | \boldsymbol{h}^{\boldsymbol{c},d}, \boldsymbol{Q}\right), \tag{4.23}$$

where $\boldsymbol{\mathcal{R}}$ represents the features extracted by a filter bank $\mathcal{F}$, and $\boldsymbol{Q}$ denotes any information other than the PSF, which may be related to all-in-focused image or noise.

Specifically, Zhu et al. [62] employed a Gabor filter bank $\mathcal{F} = \{\boldsymbol{t}_i\}$ to extract features of the derivative of an image $\boldsymbol{g}^{\nabla}_{\boldsymbol{M}}$ locally, and the extracted features can be denoted by $\left\{\boldsymbol{g}^{\nabla}_{\boldsymbol{M}_i}\right\}$, where

$$\boldsymbol{t}_i[\boldsymbol{m}] = \boldsymbol{n}[\boldsymbol{m}] \exp\left(-j2\pi \boldsymbol{m} \boldsymbol{\xi}_i^T\right) \tag{4.24}$$

$$\boldsymbol{g}^{\nabla}_{\boldsymbol{M}_i}[\boldsymbol{m}] = \boldsymbol{g}^{\nabla}_{\boldsymbol{M}}[\boldsymbol{m}] \otimes \boldsymbol{t}_i[\boldsymbol{m}], \tag{4.25}$$

where $\boldsymbol{n}[\boldsymbol{m}]$ is a 2D Gaussian function. Then the likelihood distribution of the

extracted features of $\boldsymbol{g}_{\boldsymbol{M}}^{\nabla}$ is modelled as

$$p\left(\boldsymbol{\mathcal{R}}|\boldsymbol{h}^{\boldsymbol{c},d},\boldsymbol{Q}\right) = p\left(\left\{|\boldsymbol{g}_{\boldsymbol{M}_i}^{\nabla}[\boldsymbol{m}]|^2\right\}|\boldsymbol{h}^{\boldsymbol{c},d},s\right) = \prod_i \boldsymbol{Exp}\left(|\boldsymbol{g}_{\boldsymbol{M}_i}^{\nabla}[\boldsymbol{m}]|^2; \frac{1}{s\sigma_{h,j}^2+\sigma_{\omega,i}^2}\right), \tag{4.26}$$

where $\boldsymbol{Exp}$ is the exponential distribution, s is the local variance of the derivative of all-in-focus image $\boldsymbol{f}_{\boldsymbol{M}}^{\nabla}$, since $\boldsymbol{f}_{\boldsymbol{M}}^{\nabla} \sim \mathcal{N}(0,s)$ is assumed; $\left\{\sigma_{h,i}^2\right\}$ and $\left\{\sigma_{\omega,i}^2\right\}$ are extracted spectrum of the PSF and noise, respectively, defined as

$$\sigma_{h,i}^2 = \|\boldsymbol{h} \otimes \boldsymbol{t}_i\|_2^2 \tag{4.27}$$

$$\sigma_{\omega,i}^2 = \sigma_{\omega}^2 \|\nabla \boldsymbol{t}_i\|_2^2, \tag{4.28}$$

where σ_{ω}^2 is the variance of Gaussian noise, and ∇ is the derivative operator. Since s is unknown due to the lack of prior information, it is generally estimated by maximising the likelihood given in Eq. (4.26) when $\boldsymbol{h}$ is fixed [62]. That is,

$$p\left(\left\{|\boldsymbol{g}_{\boldsymbol{M}_i}^{\nabla}[\boldsymbol{m}]|^2\right\}|\boldsymbol{h}^{\boldsymbol{c},d}\right) \propto p\left(\left\{|\boldsymbol{g}_{\boldsymbol{M}_i}^{\nabla}[\boldsymbol{m}]|^2\right\}|\boldsymbol{h}^{\boldsymbol{c},d},\hat{s}_h\right), \tag{4.29}$$

where

$$\hat{s}_h = \arg\max_s p\left(\left\{|\boldsymbol{g}_{\boldsymbol{M}_i}^{\nabla}[\boldsymbol{m}]|^2\right\}|\boldsymbol{h}^{\boldsymbol{c},d},s\right). \tag{4.30}$$

On the other hand, instead of using a large filter bank to extract most of frequency components, Burge and Geisler [6] employed a statistical learning algorithm, which is known as accuracy maximising analysis (AMA) [13], to learn an optimal filter bank, which extracts only a few of key spatial frequency features for distinguishing different depth, from training images. That is, AMA does dimensionality reduction. Once the filter bank $\mathcal{F}$ is determined, it is applied to a training set containing images blurred by the same PSF to learn the corresponding likelihood function, which is fitted to a multivariate Gaussian distribution, as

$$p\left(\boldsymbol{\mathcal{R}}|\boldsymbol{h}^{\boldsymbol{c},d},\boldsymbol{Q}\right) = p\left(\left\{|\mathcal{F}\boldsymbol{g}_{\boldsymbol{M}}|^2\right\}|\boldsymbol{h}^{\boldsymbol{c},d}\right) \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}), \tag{4.31}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and covariance matrix of the feature vectors of training images, respectively [6].

The procedure of depth estimation again contains two parts, and it is done patch-wisely with PSFs pre-sampled at a finite set of depths $\mathcal{K}$. In the first part, for each $d_k \in \mathcal{K}$, a filter bank $\mathcal{F}_{d_k}$ for feature extraction is either calculated by using Eq. (4.24) or learnt from a training set via AMA.

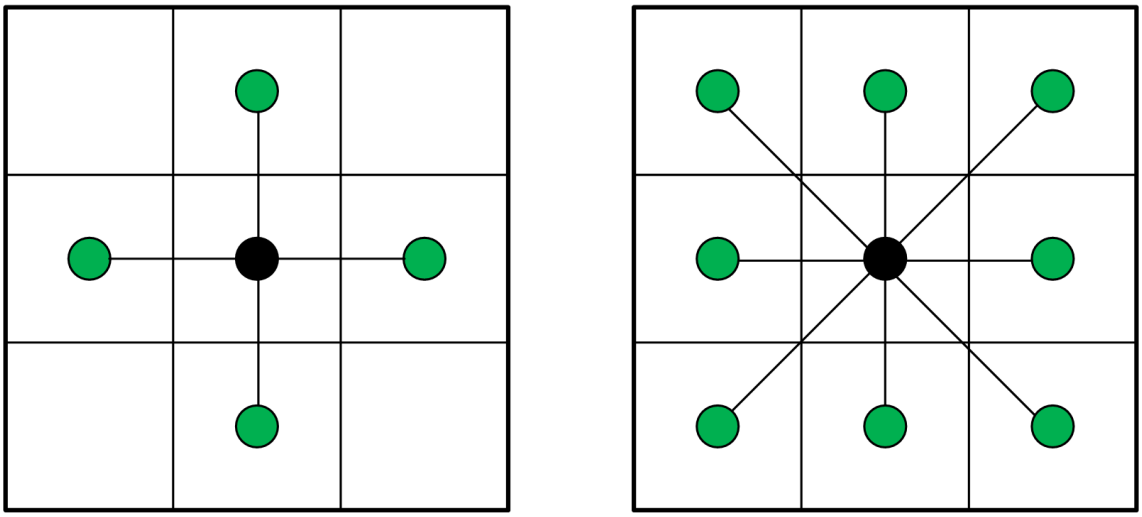


Figure 4.2 *Left: Illustration of a $\mathcal{N}_4$ neighbourhood. Right: Illustration of a $\mathcal{N}_8$ neighbourhood.*

The second part is MLE. The filter bank is applied to each image patch $\boldsymbol{g_L}$, and extracted feature spectra are denoted by $\boldsymbol{\mathcal{R}_{g_L}}$. Depth estimation is done by finding the PSF that maximises the likelihood function, as follows

$$\boldsymbol{D_M}^*[l] = \arg\max_{d_k} p\left(\boldsymbol{\mathcal{R}_{g_L}} | \boldsymbol{h}^{\boldsymbol{c},d_k}\right). \tag{4.32}$$

The procedure is repeated for all pixels.

4.4 Depth map post-processing

Aforementioned methods are all based on the local space-invariant assumption and solve depth estimation locally. In these cases, the resulting depth map $\boldsymbol{D_M}$ is considered as a raw depth map.

An improvement can be achieved by employing Markov random field (MRF) to post-process the raw depth map. A MRF is analysed in the Bayesian framework by introducing a smoothness prior to depth map. When the restoration-based strategy is used, it is

$$\boldsymbol{D_M}^*, \boldsymbol{f_M}^* = \arg\max_{\boldsymbol{D_M}, \boldsymbol{f_M}} \prod_{n=1}^{N} \left\{ p\left(\boldsymbol{g}_{\boldsymbol{M}_n} | \boldsymbol{H}_{\boldsymbol{C}_{\boldsymbol{M}_n}, \boldsymbol{D_M}} \boldsymbol{f_M}\right) \right\} p\left(\boldsymbol{D_M}\right) p\left(\boldsymbol{f_M}\right), \tag{4.33}$$

while for methods of the restoration-free strategy, it becomes

$$\boldsymbol{D_M}^* = \arg\max_{\boldsymbol{D_M}} \prod_{n=1}^{N} \left\{ p\left(\left\{ \mathcal{F} \boldsymbol{g_{M}}_n \right\} | \boldsymbol{H_{C_{M}}}_n, \boldsymbol{D_M}, \boldsymbol{Q} \right) \right\} p\left(\boldsymbol{D_M} \right). \tag{4.34}$$

In both cases, the prior distribution $p\left(\boldsymbol{D_M}\right)$ is usually a Gibbs distribution, which is in the exponential family. Particularly, the negative log of $p\left(\boldsymbol{D_M}\right)$ can be represented as

$$-\log p\left(\boldsymbol{D_M}\right) = \sum_{(i,j)\in\mathcal{N}} V_{i,j}\left(d_i, d_j\right), \tag{4.35}$$

where V is a potential function and (i, j) denotes a pair of neighbouring pixels [5]. Commonly the neighbourhood $\mathcal{N}$ used for depth estimation are $\mathcal{N}_4$ and $\mathcal{N}_8$, and they mean that the estimated depth at a particular pixel is influenced by its four direct neighbours ($\mathcal{N}_4$) or additionally four diagonal neighbours ($\mathcal{N}_8$), as shown in Figure 4.2. Finally, Eq. (4.33) or Eq. (4.34) can be solved by using e.g. Graph cuts [5].

5. CODED APERTURE: REVIEW AND DEVELOPMENT

In this chapter, the key technique of this thesis, referred to as coded aperture, is introduced as a tool improving the performance of DfD. Traditionally, DfD uses images captured by off-the-shelf cameras, whose apertures are of circular or hexagonal shape. However, the coded aperture approach uses cameras with masks on their aperture positions. The masks follow specific codes, therefore, the term "coded aperture" is used. For coded aperture, understanding its principle and designing optimised mask patterns are two aspects of importance. This chapter starts with motivating the use of coded masks. Then it presents its principle of operation and discusses two strategies of designing optimised mask patterns.

5.1 PSF modification

As shown in Chapter 4, the DfD is done by comparing candidate PSFs with the PSF encoded in images, where similarity is measured by certain criteria, e.g. some energy functions. A good energy function is expected to be minimised when two PSFs are the same, while high values are given if they are different, and the value is proportional to the amount of difference. However, for existing energy functions, two PSFs derived from a conventional aperture cause a small difference, even when two PSFs differ considerably. Also, the corresponding subspaces of those PSFs are largely overlapping. Further, a conventional aperture has a symmetrical shape, which leads to similar blurring effect on different sides of the focused distance. So when only a single image is available, distinguishing sides is challenging, and this is known as sign problem [45]. Those problems not only limit the depth resolution of DfD, but also make it prone to noise.

Intuitively, the reason causing this insignificant difference is that the defocus PSF derived from a conventional aperture works like a low-pass filter, which attenuates high frequency components heavily. Therefore, the comparison is done only with remained low frequency components. In addition, the lack of high frequency components makes image restoration, if required, considerably difficult. That is, the

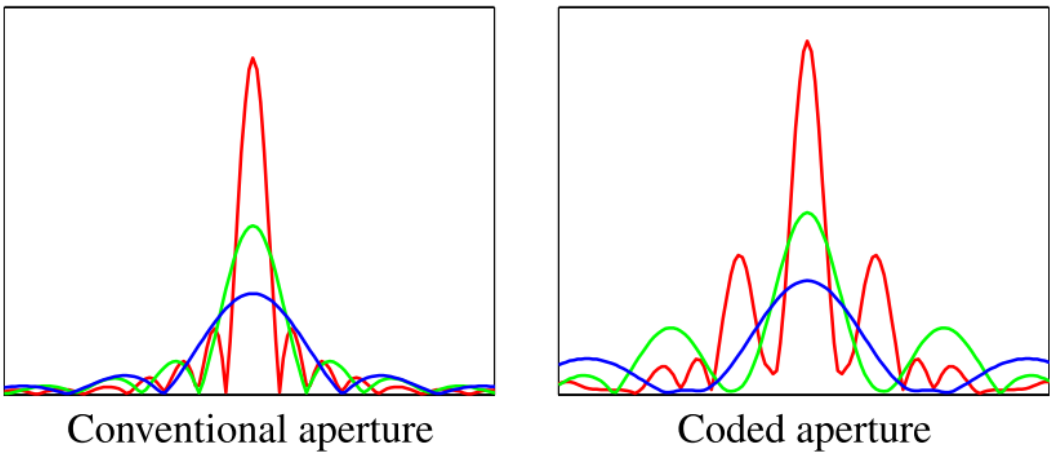


Figure 5.1 *The Fourier transforms of PSFs from conventional and coded aperture at three different scales in 1D case [23]. Reprinted by permission. ©2007 Association for Computing Machinery, Inc.*

blurring effect of a conventional PSF is undesired [17]. However, two major questions remain untouched: what is the desired blurring effect and which kind of mask pattern achieves it?

Regarding the first question, the most important works were done by Dowski and Cathey [8], [9]. Their significant contribution is a necessary condition for single image depth estimation via defocus blur cue, which says that the frequency responses of a PSF must have regions of zeros in its frequency response. This necessary condition reveals that the depth information lies in the zero-crossings of the frequency responses of the PSF, and the positions of zero-crossings need to be changing with the depth. They also experimentally noticed that it would be better if the zero-crossings appear periodically and if the PSF's frequency responses are as high as possible at non-zero parts, especially in the vicinity of zero-crossings [8]. By taking into account the noise effect, Levin et al. [23] further pointed out that for the depth estimation purpose, it would be better if zero-crossings are at low frequencies. This is because most of images' energies are concentrated at low frequencies, so zero-crossings at these regions will be clearly distinguishable and thus be more robust to noise. On the other hand, during studying wave-front coding, Dowski and Cathey [9] pointed out that for image restoration, it is preferred to modify the frequency responses of a PSF in such a way that it has a wide passband without zero-crossings, since information will be lost at zero-crossings. Therefore, depth estimation and image restoration set different, even contradictory, preference on the frequency response of PSF. However, as illustrated on the left of Figure 5.1, where a 1D slice of the frequency responses of three PSFs derived from a conventional aperture at dif-

ferent depths are compared, it is obvious that zero-crossings are unclear and largely overlapping rather than being depth dependent and the passbands are all narrow. That is, a conventional aperture satisfies neither of the requirements well, and this observation supports the intuitive opinion. The design of desired PSFs can be done by inserting a mask with certain pattern on the position of the aperture. When a coded mask is inserted, the frequency responses of its PSF is modified accordingly. As a comparison to the conventional aperture case, a similar figure for the case of a coded aperture for the depth estimation purpose is given in the right of Figure 5.1, where zero-crossings are more clear and less overlapping, which makes the depth discrimination easier.

The above discussion gives a clear direction for the mask pattern design. It also indicates that different tasks need different, even conflicting, modifications. Therefore, generally there is no universal optimised pattern for all purposes, but for a special purpose, a task specific mask pattern can be optimised to achieve the desired modification. For this reason, Hiura and Matsuyama made an insightful statement that designing desired blurring effect is the essence of DfD [17].

5.2 Masks pattern design: early examples

Before introducing the general framework for mask pattern design, two pioneering works on the mask design for the camera imaging system are presented.

To our knowledge, the first important mask design work for the camera imaging system was reported by Dowski and Cathey [8] for the depth estimation purpose. In their work, they found a simple mask leading to PSFs with depth dependent zero-crossings in the frequency domain. However, the method they used to find such a simple mask can only produce masks with low optical efficiency ($\sim 4\%$), which means that most of the light is blocked by the mask. This problem was solved by summing up this simple mask, known as the reference mask, with multiple dual masks. Those dual masks are of high optical efficiency, and adding them has little effects on the positions of zero-crossings of PSFs derived from the reference mask. Finally, they obtained a mask with 32% optical efficiency.

Another important work was done by Farid and Simoncelli [11], who designed a pair of optical masks for depth estimation. One mask is of a Gaussian pattern $M(u)$; the pattern of the other mask is the derivative of this Gaussian pattern $M'(u)$, performing as a differential operator on the image. However, this pair is not directly usable since the differentiation mask contains negative values. An appropriate mask pair was obtained by a combination of the original masks.

These two examples together reveal two options of the mask design. One option is that the mask could be of either binary or continuous values. The advantage of a binary mask is its easy manufacture, since the accurate constructions of masks of continuous values are practically difficult [23]. However, a continuous mask offers more degrees of freedom for the pattern design over a binary mask. The other option is that one can design either a single mask or a pair of (or multiple) masks. For a single mask design, generally its optical efficiency should be taken into consideration, while for a pair of masks, the relationship between two masks is of prime attention.

5.3 Masks pattern design: brute force search

In this section, as a standard problem-solving strategy, the brute force search is utilised to design an optimised single mask or an optimised pair of masks, which can enhance the performance of DfD and/or image restoration.

There are two stages in the mask design, namely generation and testing. As mentioned in Section 3.3, an aperture pattern can be considered a combination of several basic patterns. In the generation stage, for simplicity, it is common to view the full aperture A as a square rather than a circle. In addition, this square aperture can be assumed to be formed by an $n \times n$ uniform grid of small squares, and those small squares are considered as elementary apertures A_i. If we further assume that within the area of each elementary aperture, the aperture has an uniform transmission efficiency a_i, then we have

$$A = \sum_{i=1}^{n^2} a_i A_i. \tag{5.1}$$

More importantly, when n is fixed, $\{a_i\}$ gives a complete representation of an aperture pattern, thus designing a mask pattern is equivalent to determining the coefficients $\{a_i\}$.

There are two considerations about Eq. (5.1). One is about the number of basic squares n. Obviously, when n is large, a fine mask pattern can be expected. However, using a large n causes both computational and physical problems, since it increases the number of variables quadratically and decreases the size of each elementary aperture. When the size of each elementary aperture is too small, the influence of diffraction would be heavy, which makes applying the simple geometrical optics model unreasonable. Therefore, n should be chosen properly. For example, Levin et al. set $n = 13$, corresponding to a 1 mm^2 block in their case [23]. The

other consideration is about the range of each a_i. Generally, as a coefficient representing transmission efficiency, a_i could take any value between 0 and 1, from completely blocking to completely transmitting. However, it makes the searching space infinitely large and intractable. Therefore, when the brute force strategy is employed, generally the value of a_i is restricted to be either 0 or 1, which means either close or open, and it leads to a binary mask.

Taking into consideration those practical issues mentioned above, there are 2^{n^2} possibilities in total for a single binary mask case, while for the case of a pair of binary masks, the number becomes 4^{n^2}. In order to find the optimised mask pattern, usually a large set of binary masks is randomly sampled as candidates. Additional constraints may be applied to this sampled set to eliminate unwanted candidates. For example, from a manufacturing point of view, the pattern should lead to a complete mask without unconnected floating parts [23]; or the pattern should have a sufficiently large optical efficiency [30]. Then, for each valid candidate, a set of PSFs are derived from it at a set of depths $\mathcal{K}$.

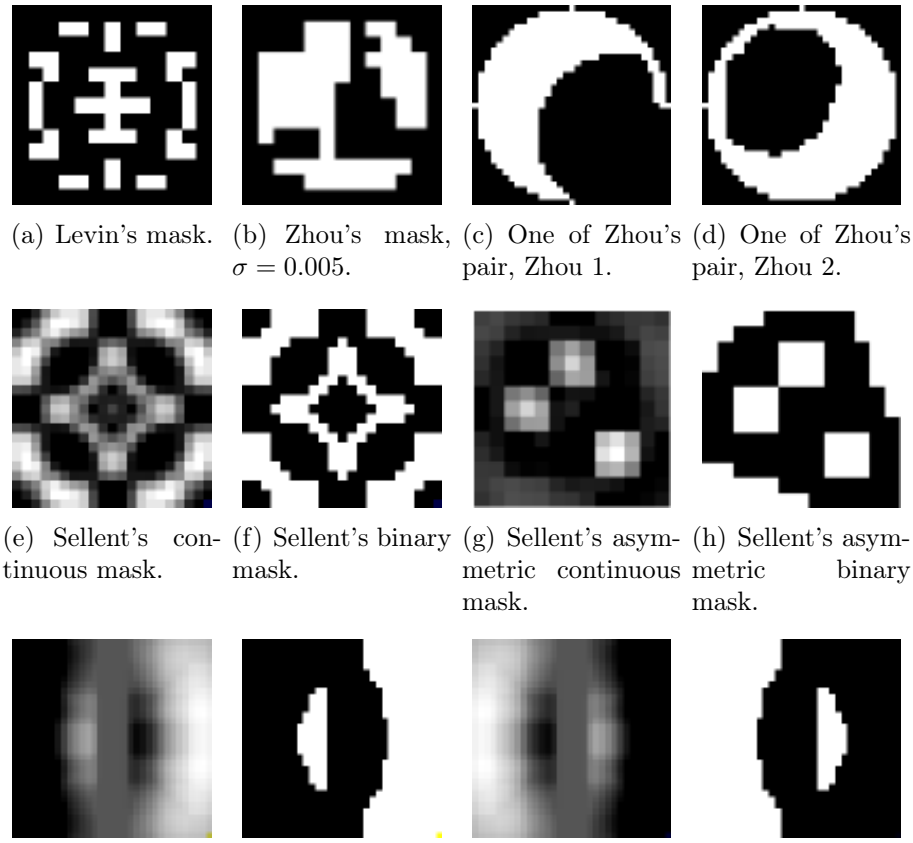
In the testing stage, those valid candidates are evaluated. In order to do evaluation, proper criteria should be defined in accordance with the task, and quite often criteria are defined based on a certain kind of solving strategy.

For designing a single mask for DfD, Levin et al. [23] proposed a depth discrimination criterion. As mentioned in Section 4.2, locally the depth is estimated by selecting the PSF $\boldsymbol{h}^{d_k}$ that maximises the likelihood function $p(\boldsymbol{g_M}|\boldsymbol{h}^{d_k})$ from pre-sampled PSFs derived from the same aperture. Intuitively, $p(\boldsymbol{g_M}|\boldsymbol{h}^{d_{k_i}})$ and $p(\boldsymbol{g_M}|\boldsymbol{h}^{d_{k_j}})$ should be sufficiently different to distinguish any pair of PSFs. Based on this intuition, the Kullback-Leibler(KL) divergence is used to measure the difference, as follows,

$$\begin{aligned} &D_{KL}\left(p\left(\boldsymbol{g_M}|\boldsymbol{h}^{d_{k_i}}\right), p\left(\boldsymbol{g_M}|\boldsymbol{h}^{d_{k_j}}\right)\right) = \\ &\int_{\boldsymbol{g_M}} p\left(\boldsymbol{g_M}|\boldsymbol{h}^{d_{k_i}}\right)\left(log\left(p\left(\boldsymbol{g_M}|\boldsymbol{h}^{d_{k_i}}\right)\right) - log\left(p\left(\boldsymbol{g_M}|\boldsymbol{h}^{d_{k_j}}\right)\right)\right)\mu\left(\boldsymbol{g_M}\right). \end{aligned} \tag{5.2}$$

For each mask candidate, the KL divergence given in Eq. (5.2) is calculated for all pairs of corresponding PSFs at depths $\mathcal{K}$, and the minimum value is set as the score of this mask candidate. Finally, the mask candidate getting the highest value is chosen as the optimised single mask for DfD purpose. For the case of $n = 13$, the optimised mask is shown in Figure 5.2(a).

The problem of designing a single mask for image restoration is studied by Zhou and Nayar [61]. In this case, the criterion is to minimise the expectation of the L2



(a) Levin's mask. (b) Zhou's mask, $\sigma = 0.005$. (c) One of Zhou's pair, Zhou 1. (d) One of Zhou's pair, Zhou 2.

(e) Sellent's continuous mask. (f) Sellent's binary mask. (g) Sellent's asymmetric continuous mask. (h) Sellent's asymmetric binary mask.

(i) One of Sellent's continuous pair. (j) One of Sellent's binary pair. (k) One of Sellent's continuous pair. (l) One of Sellent's binary pair.

Figure 5.2 *Examples of optimised mask patterns.*

Table 5.1 *Genetic algorithm for aperture pattern optimisation (adapted from Table 1 in [61]). Reprinted by permission. ©2009 IEEE*

STEPS:	
1 :	Initialise: $g = 0$; randomly generate S binary sequences of length n^2.
2 :	For $g = 1 : G$
2a :	Selection: For each sequence h, the corresponding H in the frequency domain is calculated and then evaluated by using Eq. (5.5). Only the best M out of S sequences are selected.
2b :	Repeat until the number of sequences increase from M to S. -Crossover: Duplicate two randomly chosen sequences from the M sequences of Step 2a, align them, and exchange each pair of corresponding bits with a probability of p_1, to obtain two new sequences. -Mutation: For each newly generated sequence, flip each bit with a probability p_2.
3 :	End for
4 :	Evaluate all the remaining sequences using Eq. (5.5) and output the best one.

norm of the difference between the DFT of a restored image $\hat{\boldsymbol{F}}_{\boldsymbol{M}}$ and the DFT of the ground truth $\boldsymbol{F}_{\boldsymbol{M}}$ in the frequency domain, as follows,

$$R\left(\boldsymbol{H}, \boldsymbol{F}_{\boldsymbol{M}}^{\mathbf{0}}, \boldsymbol{W}\right) = \mathbb{E}\left\|\hat{\boldsymbol{F}}_{\boldsymbol{M}} - \boldsymbol{F}_{\boldsymbol{M}}^{\mathbf{0}}\right\|_2^2, \tag{5.3}$$

where $\boldsymbol{W}$ is the same as defined in Eq. (4.14). By assuming a White Gaussian noise with distribution $\boldsymbol{\omega}_{\boldsymbol{M}} \sim \mathcal{N}\left(0, \sigma^2 \boldsymbol{I}\right)$ and using Eq. (3.15) and Eq. (4.16), we have

$$\begin{aligned} R\left(\boldsymbol{H}_d, \boldsymbol{F}_{\boldsymbol{M}}^{\mathbf{0}}, \boldsymbol{W}\right) &= \mathbb{E}\left\|\hat{\boldsymbol{F}}_{\boldsymbol{M}} - \boldsymbol{F}_{\boldsymbol{M}}^{\mathbf{0}}\right\|_2^2 \\ &= \mathbb{E}\left\|\frac{\boldsymbol{\Omega} \bullet \boldsymbol{H} - \boldsymbol{F}_{\boldsymbol{M}}^{\mathbf{0}} \bullet |\boldsymbol{W}|^2}{|\boldsymbol{H}|^2 + |\boldsymbol{W}|^2}\right\|_2^2 \\ &= \left\|\frac{\sigma \boldsymbol{H}}{|\boldsymbol{H}|^2 + |\boldsymbol{W}|^2}\right\|_2^2 + \left\|\frac{\boldsymbol{F}_{\boldsymbol{M}}^{\mathbf{0}} \bullet |\boldsymbol{W}|^2}{|\boldsymbol{H}|^2 + |\boldsymbol{W}|^2}\right\|_2^2. \end{aligned} \tag{5.4}$$

Furthermore, $\boldsymbol{F}_{\boldsymbol{M}}^{\mathbf{0}}$ is the DFT of a natural image, so it can be integrated out by using $\frac{1}{\boldsymbol{\xi}}$ law stated in Section 4.2, then we have

$$\begin{aligned} R\left(\boldsymbol{H}\right) &= \int_{\boldsymbol{F}_{\boldsymbol{M}}^{\mathbf{0}}} R\left(\boldsymbol{H}, \boldsymbol{F}_{\boldsymbol{M}}^{\mathbf{0}}, \boldsymbol{W}\right) \mu\left(\boldsymbol{F}_{\boldsymbol{M}}^{\mathbf{0}}\right) \\ &= \sum_{\boldsymbol{\xi}} \frac{\sigma^2}{|\boldsymbol{H}^{\mathbf{c},d}\left(\boldsymbol{\xi}\right)|^2 + |\sigma \boldsymbol{\Psi}\left(\boldsymbol{\xi}\right)|^2}, \end{aligned} \tag{5.5}$$

which shows that the noise level σ plays an important role in evaluating mask patterns [61]. That is, for different noise levels, the optimised single mask pattern for image restoration is different. For a fix noise level, a genetic algorithm is applied to find the best mask pattern minimising Eq. (5.5), see Table 5.1 for detailed steps. The optimised mask patterns for noise with $\sigma = 0.005$ is shown in Figure 5.2(b) as an example. Similar procedures have also been applied by Masia et al. [30] to find optimised single image restoration mask patterns with other criteria.

As discussed in Section 5.1, DfD and image restoration set contradictory requirements in PSFs, and thus a single mask cannot satisfy both of them simultaneously. For this reason, Zhou et al. [59] studied the problem of designing a pair of complementary masks that satisfy both requirements. Based on the Bayesian analysis done

in Section 4.2, we have

$$\begin{aligned} d^*, \boldsymbol{F_M}^* &= \arg\max_{d,\boldsymbol{F_M}} p\left(d, \boldsymbol{F_M}|\boldsymbol{G_{M_1}}, \boldsymbol{G_{M_2}}\right) \\ &= \arg\min_{d,\boldsymbol{F_M}} \sum_{n=1}^{N} \left(\left\| \boldsymbol{G_{M_n}} - \boldsymbol{H}^{\boldsymbol{c}_n,d} \bullet \boldsymbol{F_M} \right\|_2^2 \right) + \left\| \boldsymbol{W} \bullet \boldsymbol{F_M} \right\|_2^2 \\ &= \arg\min_{d,\boldsymbol{F_M}} E\left(d, \boldsymbol{F_M}|\boldsymbol{G_{M_1}}, \boldsymbol{G_{M_2}}\right). \end{aligned} \tag{5.6}$$

Utilising Eq. (3.15) and Eq. (4.16), the ground truth all-in-focus image can be represented as

$$\boldsymbol{F_M^0} = \frac{\boldsymbol{G_{M_1}} \bullet \bar{\boldsymbol{H}}^{\boldsymbol{c}_1,d_{GT}} + \boldsymbol{G_{M_2}} \bullet \bar{\boldsymbol{H}}^{\boldsymbol{c}_2,d_{GT}}}{|\boldsymbol{H}^{\boldsymbol{c}_1,d_{GT}}|^2 + |\boldsymbol{H}^{\boldsymbol{c}_2,d_{GT}}|^2 + |\boldsymbol{W}|^2}. \tag{5.7}$$

Then, a new energy function depending only on PSFs can be obtained by substituting Eq. (5.7) into Eq. (5.6) and integrating out $\boldsymbol{F_M^0}$ according to $\frac{1}{\boldsymbol{\xi}}$ law, as follows,

$$\begin{aligned} &E\left(d|\boldsymbol{H}^{\boldsymbol{c}_1,d_{GT}}, \boldsymbol{H}^{\boldsymbol{c}_2,d_{GT}}, \sigma\right) = \int_{\boldsymbol{F_M^0}} E\left(d|\boldsymbol{G_{M_1}}, \boldsymbol{G_{M_2}}\right) \mu\left(\boldsymbol{F_M^0}\right) \\ &= \int_{\boldsymbol{F_M^0}} E\left(d|\boldsymbol{F_M^0}, \boldsymbol{H}^{\boldsymbol{c}_1,d_{GT}}, \boldsymbol{H}^{\boldsymbol{c}_2,d_{GT}}, \sigma\right) \mu\left(\boldsymbol{F_M^0}\right) \\ &= \sum_{\boldsymbol{\xi}} \frac{1}{|\boldsymbol{\Psi}|^2} \frac{|\boldsymbol{H}^{\boldsymbol{c}_1,d} \bullet \boldsymbol{H}^{\boldsymbol{c}_1,d_{GT}} - \boldsymbol{H}^{\boldsymbol{c}_2,d} \bullet \boldsymbol{H}^{\boldsymbol{c}_2,d_{GT}}|^2}{|\boldsymbol{H}^{\boldsymbol{c}_1,d}|^2 + |\boldsymbol{H}^{\boldsymbol{c}_2,d}|^2 + |\boldsymbol{W}|^2} + \sigma^2 \sum_{\boldsymbol{\xi}} \left[\frac{|\boldsymbol{H}^{\boldsymbol{c}_1,d_{GT}}|^2 + |\boldsymbol{H}^{\boldsymbol{c}_2,d_{GT}}|^2 + |\boldsymbol{W}|^2}{|\boldsymbol{H}^{\boldsymbol{c}_1,d}|^2 + |\boldsymbol{H}^{\boldsymbol{c}_2,d}|^2 + |\boldsymbol{W}|^2} + n \right], \end{aligned} \tag{5.8}$$

which measures the distance between a depth d to the ground truth depth d_{GT} [60]. Then a criterion for mask pattern evaluation can be defined as

$$\begin{aligned} R\left(\boldsymbol{H}^{\boldsymbol{c}_1,d_k}, \boldsymbol{H}^{\boldsymbol{c}_2,d_k}|d_m, \sigma\right) &= \min_{d_k \in \mathcal{K}/d_m} E\left(d_k|\boldsymbol{H}^{\boldsymbol{c}_1,d_m}, \boldsymbol{H}^{\boldsymbol{c}_2,d_m}\right) - E\left(d_m|\boldsymbol{H}^{\boldsymbol{c}_1,d_m}, \boldsymbol{H}^{\boldsymbol{c}_2,d_m}\right) \\ &\approx \sum_{\boldsymbol{\xi}} \frac{1}{|\boldsymbol{\Psi}|^2} \frac{|\boldsymbol{H}^{\boldsymbol{c}_1,d_k} \bullet \boldsymbol{H}^{\boldsymbol{c}_1,d_m} - \boldsymbol{H}^{\boldsymbol{c}_2,d_k} \bullet \boldsymbol{H}^{\boldsymbol{c}_2,d_m}|^2}{|\boldsymbol{H}^{\boldsymbol{c}_1,d_k}|^2 + |\boldsymbol{H}^{\boldsymbol{c}_2,d_k}|^2 + |\boldsymbol{W}|^2}, \end{aligned} \tag{5.9}$$

where the middle depth $d_m \in \mathcal{K}$ is selected as the ground truth depth, and all other depths are compared with it. Based on this criterion, the same genetic algorithm shown in Table 5.1 can be modified and applied to find an optimised mask pair. The resulting mask pair is further refined to have higher resolution, and the counterparts of resolution 33×33 are shown in Figure 5.2(c) and Figure 5.2(d) [59].

5.4 Masks pattern design: analytic search

The brute force search has been successfully used for designing binary masks. However, when it is used to design grey-scale masks, the problem becomes intractable. In this section, an analytic search framework proposed by Sellent and Favaro [46] is introduced to solve this problem.

As mentioned in Section 4.3, PSFs at different depths define different subspaces. Therefore, for DfD, the distance between two subspaces can be used as a measure of depth discrimination [29]. In order to employ this idea for evaluating mask patterns, a sufficiently large set of natural images $\boldsymbol{F}^0_{train}$ is included, so when those images are blurred, they can be considered as spanning the whole subspace. Then, the distance between two subspaces defined by PSFs $\boldsymbol{h}^{\boldsymbol{a},d_i}$ and $\boldsymbol{h}^{\boldsymbol{a},d_j}$ can be defined as $\sum_{\boldsymbol{f}^0_{\boldsymbol{M}}\in\boldsymbol{F}^0_{train}} \left\|\mathcal{P}_{\mathcal{B}_i}\boldsymbol{g}^{\boldsymbol{a},d_i}_{\boldsymbol{M}} - \mathcal{P}_{\mathcal{B}_j}\boldsymbol{g}^{\boldsymbol{a},d_j}_{\boldsymbol{M}}\right\|^2_2$, where $\boldsymbol{a}$ is a vector representing the aperture shape, $\boldsymbol{g}^{\boldsymbol{a},d}_{\boldsymbol{M}} = \boldsymbol{f}^0_{\boldsymbol{M}} \otimes \boldsymbol{h}^{\boldsymbol{a},d}$ is an image blurred by the PSF corresponding to the depth d, and $\mathcal{P}_{\mathcal{B}}$ is the band-limiting operator as defined in Section 4.3, to eliminate the influence of image texture. For each aperture $\boldsymbol{a}$, PSFs are derived at a set of depths $\mathcal{K}$, so subspace distance is calculated pair-wise, and their summation is the object function, written as [46]

$$E(\boldsymbol{a}) = \sum_{\boldsymbol{f}^0_{\boldsymbol{M}}\in\boldsymbol{F}^0_{train}} \sum_{d_i\neq d_j} \left\|\mathcal{P}_{\mathcal{B}_i}\boldsymbol{g}^{\boldsymbol{a},d_i}_{\boldsymbol{M}} - \mathcal{P}_{\mathcal{B}_j}\boldsymbol{g}^{\boldsymbol{a},d_j}_{\boldsymbol{M}}\right\|^2_2, \forall d_i, d_j \in \mathcal{K}. \tag{5.10}$$

According to the aperture superposition principle mentioned in Section 3.3, an image blurred by a PSF at depth d can be represented as

$$\begin{aligned}\mathcal{P}_{\mathcal{B}}\boldsymbol{g}^{\boldsymbol{a},d}_{\boldsymbol{M}} &= \mathcal{P}_{\mathcal{B}}\left[\boldsymbol{g}^{1,d}_{\boldsymbol{M}}, ..., \boldsymbol{g}^{n^2,d}_{\boldsymbol{M}}\right]\boldsymbol{a} \\ &= \boldsymbol{N}^d\boldsymbol{a},\end{aligned} \tag{5.11}$$

where each column of $\boldsymbol{N}^d$ is an image vector corresponding to an elementary aperture.
Then function 5.10 can be simplified as

$$E(\boldsymbol{a}) = \boldsymbol{a}^T\boldsymbol{M}_{\boldsymbol{CA}}\boldsymbol{a}, \tag{5.12}$$

where $\boldsymbol{M}_{\boldsymbol{CA}} = \sum_{\boldsymbol{f}^0_{\boldsymbol{M}}\in\boldsymbol{F}^0_{train}} \sum_{d_i\neq d_j} \left(\boldsymbol{N}^{d_i}\right)^T\boldsymbol{N}^{d_i} + \left(\boldsymbol{N}^{d_j}\right)^T\boldsymbol{N}^{d_j} - 2\left(\boldsymbol{N}^{d_i}\right)^T\boldsymbol{N}^{d_j}$. Taking into account a few of practical considerations, e.g. optical efficiency, the mask

pattern design can be presented as a constraint optimisation problem,

$$\boldsymbol{a}^* = \arg\max_{\boldsymbol{a}} \left(\boldsymbol{a}^T \boldsymbol{M_{CA}} \boldsymbol{a} + \lambda \left\|\boldsymbol{a}\right\|_2^2\right), \tag{5.13}$$

with $\left\|\boldsymbol{a}\right\|_1 = 1$ and $a_i \geq 0$. The value $\boldsymbol{a}^*$ maximising Eq. (5.13) gives the optimised single mask pattern for DfD, and the one with a 21×21 resolution is shown in Figure 5.2(e) [46].

Particularly, when the depth set $\mathcal{K}$ contains depths on both sides of the focused distance, Eq. (5.13) can be used to design an asymmetrical single mask pattern that is able to solve the sign problem, and the resulting mask with a 13×13 resolution is shown in Figure 5.2(g) [45].

The function for designing a pair of mask patterns for DfD purpose can be easily extended from Eq. (5.10) and Eq. (5.13), as

$$\begin{aligned}
[\boldsymbol{a}^*; \boldsymbol{b}^*] &= \arg\max_{\boldsymbol{a},\boldsymbol{b}} \left\| \left(\boldsymbol{g}_{\boldsymbol{a}}^{d_i} - \boldsymbol{g}_{\boldsymbol{b}}^{d_i}\right) - \left(\boldsymbol{g}_{\boldsymbol{a}}^{d_j} - \boldsymbol{g}_{\boldsymbol{b}}^{d_j}\right) \right\|_2^2 + \lambda \left\|\boldsymbol{a}\right\|_2^2 + \lambda \left\|\boldsymbol{b}\right\|_2^2 \\
&= \arg\max_{\boldsymbol{a},\boldsymbol{b}} \boldsymbol{a}^T \boldsymbol{M_{CA}} \boldsymbol{a} + \boldsymbol{b}^T \boldsymbol{M_{CA}} \boldsymbol{b} - \boldsymbol{a}^T \boldsymbol{M_{CA}} \boldsymbol{b} - \boldsymbol{b}^T \boldsymbol{M_{CA}} \boldsymbol{a} + \lambda \left\|\boldsymbol{a}\right\|_2^2 + \lambda \left\|\boldsymbol{b}\right\|_2^2 \\
&= \arg\max_{[\boldsymbol{a};\boldsymbol{b}]} \left[\boldsymbol{a}; \boldsymbol{b}\right]^T \boldsymbol{M_{2CA}} \left[\boldsymbol{a}; \boldsymbol{b}\right] + \lambda \left\|\boldsymbol{a}\right\|_2^2 + \lambda \left\|\boldsymbol{b}\right\|_2^2,
\end{aligned} \tag{5.14}$$

with constraints that $\left\|\boldsymbol{a}\right\|_1 = 1$, $\left\|\boldsymbol{b}\right\|_1 = 1$, $a_i \geq 0$ and $b_i \geq 0$, where $\boldsymbol{M_{2CA}} = \begin{bmatrix} \boldsymbol{M_{CA}} & -\boldsymbol{M_{CA}} \\ -\boldsymbol{M_{CA}} & \boldsymbol{M_{CA}} \end{bmatrix}$. Notice that in Eq. (5.14), the projection operator disappears, since now the influence of image texture is eliminated by comparing two images, i.e. one is the reference of the other. The resulting $[\boldsymbol{a}^*; \boldsymbol{b}^*]$ give the optimised mask pattern pair for DfD, the pair with a 33×33 resolution is shown in Figure 5.2(i) and Figure 5.2(k) [46].

The optimised grey-scale single mask pattern and mask pattern pair can also be reduced to binary mask patterns by setting a threshold, and the resulting binary counterparts are shown in Figure 5.2(f), Figure 5.2(h), Figure 5.2(j) and Figure 5.2(l).

6. CODED APERTURE: SIMULATIONS AND EXPERIMENTS

In this chapter, three implemented DfD algorithms are tested with images from coded aperture cameras, in both simulated and real environments. The first algorithm is denoted as Levin's algorithm, which is a restoration-based algorithm requiring a single image; the second algorithm is also a restoration-based one denoted as Zhou's algorithm, which requires two images. The principle behind both Levin's algorithm and Zhou's algorithm is introduced in Section 4.2. The third algorithm is a restoration-free algorithm requiring a single image, denoted as Favaro's algorithm, as presented in Section 4.3. The step-by-step procedures of three algorithms are summarised in Table 6.1, Table 6.2 and Table 6.3, respectively.

6.1 PSF

As mentioned in Chapter 4, the success of all DfD algorithms rely on having a set of high quality PSFs. Therefore, in Section, the problem of acquiring PSFs is addressed.

Generally, PSFs can be obtained by either measurements or calculations. For measurement cases, there are two ways. A simple way is to generate a tiny point light source and put it at the sampled depth, then the intensity normalised image of this

Table 6.1 *The procedure of Levin's algorithm.*

INPUTS:	
$\boldsymbol{g_M}$:	captured image with coded aperture camera;
PSFs :	PSFs pre-sampled at a set of depths $\mathcal{K}$;
	the PSF at depth d_k is denoted as $\boldsymbol{h^{c,d_k}}$;
STEPS:	
1 :	For each PSF $\boldsymbol{h^{c,d_k}}$ at depth d_k
2 :	Obtain $\hat{\boldsymbol{f}}^k_{\boldsymbol{M}}$ by solving Eq. (4.11)
3 :	End for
4 :	Obtain depth map by solving Eq. (4.12), $\forall pixel$

Table 6.2 *The procedure of Zhou's algorithm.*

INPUTS:	
$\left(\boldsymbol{g}_{\boldsymbol{M}_1}, \boldsymbol{g}_{\boldsymbol{M}_2}\right)$:	captured images with different masks;
PSFs :	PSF pairs pre-sampled at a set of depth $\mathcal{K}$;
	the PSF pair at depth d_k is denoted as $\left(\boldsymbol{h}^{\boldsymbol{c}_1,d_k}, \boldsymbol{h}^{\boldsymbol{c}_2,d_k}\right)$;
STEPS:	
1 :	For each PSFs pair $\left(\boldsymbol{h}^{\boldsymbol{c}_1,d_k}, \boldsymbol{h}^{\boldsymbol{c}_2,d_k}\right)$ at depth d_k
2 :	Obtain $\hat{\boldsymbol{F}}^k_{\boldsymbol{M}}$ by solving Eq. (4.16)
3 :	End for
4 :	Obtain depth map by solving Eq. (4.17), $\forall pixel$

Table 6.3 *The procedure of Favaro's algorithm.*

INPUTS:	
$\boldsymbol{g}_{\boldsymbol{M}}$:	captured image with coded aperture camera;
PSFs :	PSFs pre-sampled at a set of depths $\mathcal{K}$;
	the PSF at depth d_k is denoted as $\boldsymbol{h}^{\boldsymbol{c},d_k}$;
$\boldsymbol{F}^0_{train}$:	a set of training images;
STEPS:	
1 :	For each PSF $\boldsymbol{h}^{\boldsymbol{c},d_k}$ at depth d_k
2 :	Obtain projection operator $\boldsymbol{H}^{\perp}_{\boldsymbol{c},d_k}$ by using Eq. (4.19) and Eq. (4.20)
3 :	End for
4 :	For each pixel l of the image
5 :	Take the patch L centred in l and solve Eq. (4.22)
6 :	End for

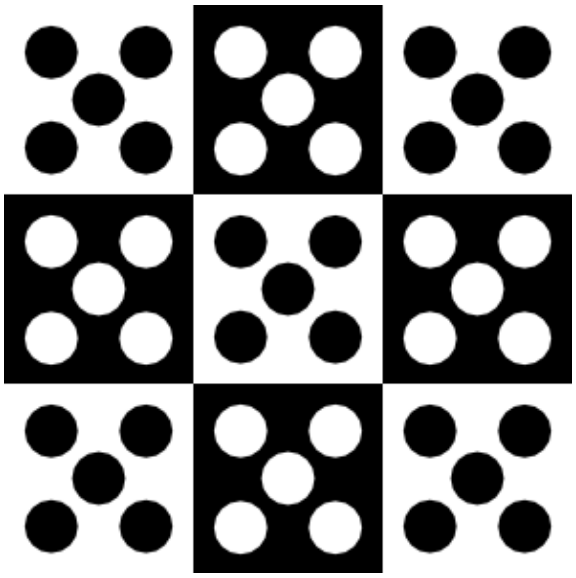


Figure 6.1 *The test pattern proposed in [19].*

point light source is considered as the PSF at that depth. However, practically creating a near ideal point light source is challenging. For example, simply drilling an opaque material and putting it in front of an uniform light source leads to a strong diffraction effect, and thus the pattern of PSFs are largely destroyed [49]. Facing this problem, a more complicate way is proposed by Joshi et al. [18] and Kee et al. [19]. In this method, instead of directly imaging a point light source, a fronto-parallel plane with a specific pattern is put at the sampled depth and imaged. As shown in Figure 6.1, an ideal pattern should contain edges of all directions, so that it can record a 2D PSF completely. Since it is a known pattern, the PSF at that depth can be obtained as

$$\boldsymbol{h}^* = \arg\min_{\boldsymbol{h}} \left\| \boldsymbol{g_M} - \boldsymbol{f}^0_{\boldsymbol{M}} \otimes \boldsymbol{h} \right\|_2^2, \tag{6.1}$$

where $\boldsymbol{g_M}$ and $\boldsymbol{f}^0_{\boldsymbol{M}}$ are both known [18]. However, although this method eventually gives accurate PSFs, it is impractical to always measure a set of PSFs together with the scene images.

Fortunately, measuring PSFs is avoidable since PSFs can also be calculated. In geometric optics case, a PSF is considered as a scaled version of aperture mask pattern, assuming aberration-free lens. Thus, a PSF at depth d can be calculated as follows: assuming that the depth d deviates from the focused distance d_f, as shown in Figure 6.2, the continuous PSF can be obtained as

$$k^d(\boldsymbol{y}) = M\left(\frac{l_d}{l_d - l_f}\boldsymbol{y}\right), \tag{6.2}$$

where $M(\boldsymbol{y})$ denotes the mask function and $l_f = \left(\frac{1}{f} - \frac{1}{d_f}\right)^{-1}$ and $l_d = \left(\frac{1}{f} - \frac{1}{d}\right)^{-1}$. Then, this continuous PSF k^d is sampled by the sensor grid as

$$\boldsymbol{h}^d[\boldsymbol{m}] = \int_{\Gamma} p_{\boldsymbol{m}}(\boldsymbol{y})\, k^d(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}, \tag{6.3}$$

and $\boldsymbol{h}^d$ is the calculated discrete PSF at depth d.

On the other hand, the wave optics enables us to model PSFs more accurately by taking into account diffraction effects. However, the derivation is more complicated. Considering temporally coherent (monochromatic) as well as spatially coherent illumination, let us start with the assumptions that we have 'an equivalent' thin lens model for our camera imaging system where we can also assume that the place of aperture and that of lens plane are coincident [3]. Furthermore, we assume that

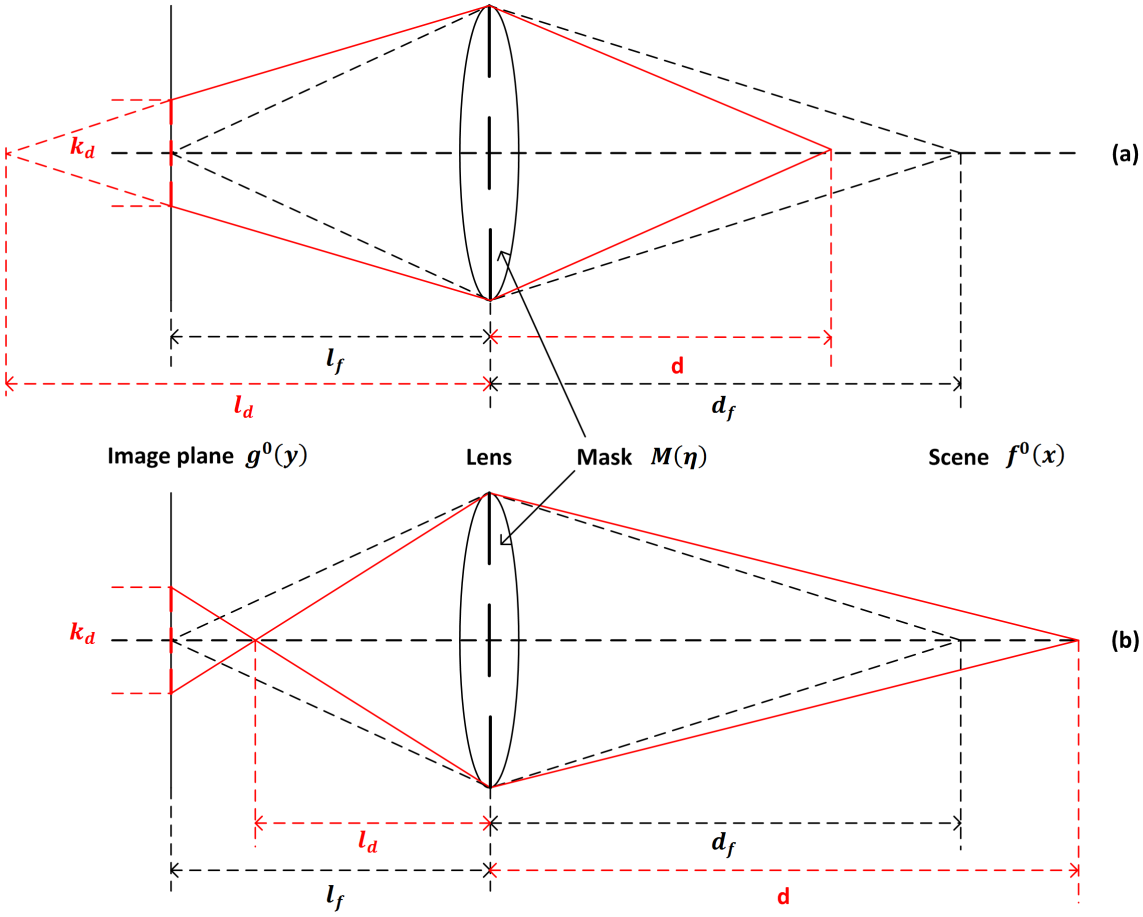


Figure 6.2 *Illustration of defocus blur in coded aperture imaging system. Top: The PSF of a point light source closer than the focused distance. Bottom: The PSF of a point light source further than the focused distance.*

we are using paraxial approximation [15] and thus utilise Fresnel diffraction model. Finally, we also assume that the lens is aberration-free.

Let the lens aperture function be defined as

$$P(\boldsymbol{\eta}) = \begin{cases} 1, & \text{inside the lens aperture;} \\ 0, & \text{otherwise.} \end{cases} \tag{6.4}$$

Following Figure 6.2, we have

$$U_M^-(\boldsymbol{\eta}) = \mathcal{FP}_d\left\{f^0(\boldsymbol{x})\right\}, \tag{6.5}$$

$$U_M^+(\boldsymbol{\eta}) = U_M^-(\boldsymbol{\eta})\, M(\boldsymbol{\eta})\, P(\boldsymbol{\eta}) \exp\left(-j\frac{\pi}{\lambda f}\left(\eta_1^2+\eta_2^2\right)\right), \tag{6.6}$$

$$g^0(\boldsymbol{y}) = \mathcal{FP}_{l_f}\left\{U_M^+(\boldsymbol{\eta})\right\}, \tag{6.7}$$

where λ is the wavelength of monochromatic light, f is the focal length of the lens, $M(\boldsymbol{\eta})$ is the mask function, and $\mathcal{FP}_z\{U(\boldsymbol{x})\}$ denotes the Fresnel propagation of $U(\boldsymbol{x})$ by distance z, given as

$$\begin{aligned}\mathcal{FP}_z\{U(\boldsymbol{x})\} &\triangleq U_z(\boldsymbol{y}) \\ &= \frac{\exp\left(j\frac{2\pi z}{\lambda}\right)}{j\lambda z}\iint\limits_{\mathbb{R}^2} U(\boldsymbol{x})\exp\left\{j\frac{\pi}{\lambda z}[(y_1-x_1)^2+(y_2-x_2)^2]\right\}\mathrm{d}x_1\mathrm{d}x_2.\end{aligned} \tag{6.8}$$

In addition, $U_M^-(\boldsymbol{\eta})$ and $U_M^+(\boldsymbol{\eta})$ are the wave fields just before and just after the lens plane, respectively. Hence, for a point light source $\boldsymbol{p_0} = [\boldsymbol{x_0}, d]^\top$, we have the coherent impulse response k_{coh} as [15]

$$\begin{aligned}k_{coh}^d(\boldsymbol{y};\boldsymbol{x_0}) =& A(\boldsymbol{y},\boldsymbol{x_0})\iint\limits_{\mathbb{R}^2}\tilde{P}(\boldsymbol{\eta})\exp\left\{j\frac{\pi z_d(\eta_1^2+\eta_2^2)}{\lambda}\right\} \\ &\exp\left\{-j\frac{2\pi}{\lambda l_f}[(y_1-\alpha x_{0_1})\eta_1+(y_2-\alpha x_{0_2})\eta_2]\right\}\mathrm{d}\eta_1\mathrm{d}\eta_2,\end{aligned} \tag{6.9}$$

where

$$\tilde{P}(\boldsymbol{\eta}) = M(\boldsymbol{\eta})P(\boldsymbol{\eta}), \tag{6.10}$$

$$z_d = \frac{1}{d}+\frac{1}{l_f}-\frac{1}{f}, \tag{6.11}$$

$$\alpha = -\frac{l_f}{d}, \tag{6.12}$$

$$A(\boldsymbol{y},\boldsymbol{x_0}) = \frac{\exp\left\{j\frac{2\pi}{\lambda}(d+l_f)\right\}}{\lambda^2 d l_f}\exp\left\{j\frac{\pi}{\lambda l_f}\left(y_1^2+y_2^2\right)\right\}\exp\left\{j\frac{\pi}{\lambda d}\left(x_{0_1}^2+x_{0_2}^2\right)\right\}. \tag{6.13}$$

Note that the imaging system is shift-invariant for the scaled scene coordinates $(\tilde{x}_1,\tilde{x}_2) = (\alpha x_1, \alpha x_2)$, i.e. k_{coh} is a function of $(y_1-\tilde{x}_1, y_2-\tilde{x}_2)$.

If the illumination is perfectly spatially incoherent, but still monochromatic, the imaging system behaves linearly for intensity rather than amplitude, and in this case, the incoherent impulse responses k_{inc} is given in terms of the coherent PSF

as [15], [55]

$$k_{inc}^{d}\left(y_1-\tilde{x}_1,y_2-\tilde{x}_2\right)=\left|k_{coh}^{d}\left(y_1-\tilde{x}_1,y_2-\tilde{x}_2\right)\right|^2$$
$$=\left|\frac{1}{\lambda^2 dl_f}\iint\limits_{\mathbb{R}^2}\tilde{P}\left(\boldsymbol{\eta}\right)\exp\left\{j\frac{\pi z_d\left(\eta_1^2+\eta_2^2\right)}{\lambda}\right\}\exp\left\{-j\frac{2\pi}{\lambda l_f}\left[\left(y_1-\tilde{x}_1\right)\eta_1+\left(y_2-\tilde{x}_2\right)\eta_2\right]\right\}\mathrm{d}\eta_1\mathrm{d}\eta_2\right|^2 \tag{6.14}$$

The k_{inc}^{d} obtained for the monochromatic case can be further generalised to polychromatic illumination, by taking into account all the desired spectral components Λ. If the imaging for the monochromatic and spatially incoherent case is given as

$$g^0\left(\boldsymbol{y}\right)=\iint\limits_{\mathbb{R}^2}f^0\left(\tilde{\boldsymbol{x}};\lambda_0\right)k_{inc}^{d}\left(y_1-\tilde{x}_1,y_2-\tilde{x}_2,\lambda_0\right)\mathrm{d}\tilde{x}_1\mathrm{d}\tilde{x}_2, \tag{6.15}$$

then for the polychromatic case, it is

$$g^0\left(\boldsymbol{y}\right)=\iint\limits_{\mathbb{R}^2}\int\limits_{\Lambda}f^0\left(\tilde{\boldsymbol{x}};\lambda\right)k_{inc}^{d}\left(y_1-\tilde{x}_1,y_2-\tilde{x}_2,\lambda\right)\mathrm{d}\lambda\mathrm{d}\tilde{x}_1\mathrm{d}\tilde{x}_2. \tag{6.16}$$

From the imaging system point of view, a weighting can be applied to PSFs for different λ's so that a colour component with a particular spectral distribution can be found as

$$g^0\left(\boldsymbol{y}\right)=\iint\limits_{\mathbb{R}^2}\int\limits_{\Lambda}f^0\left(\tilde{\boldsymbol{x}};\lambda\right)k_{inc}^{d}\left(y_1-\tilde{x}_1,y_2-\tilde{x}_2,\lambda\right)\boldsymbol{W}\left(\lambda\right)\mathrm{d}\lambda\mathrm{d}\tilde{x}_1\mathrm{d}\tilde{x}_2, \tag{6.17}$$

where $\boldsymbol{W}\left(\lambda\right)$ represents the spectral distribution of e.g. green colour for the sensor detecting 'green component' of the incident light. In other words, it is the spectral sensitivity for this particular sensor.

Some calculated PSFs with Levin's mask are shown in Figure 6.3 as examples. Those examples show that two methods lead to PSFs with different appearances, especially when the scales of PSFs are not large. The reason is that the geometrical optics is unrealistic in those areas and thus cannot produce accurate results. As the PSF scale increases, the differences between the geometrical optics and the wave optics becomes less significant, and that is the reason why PSFs obtained by two methods become similar in large scale cases.

As required by algorithms introduced in Chapter 4, PSFs at a set of depths $\mathcal{K}$ should be either measured or calculated in advance. Ideally, depths in $\mathcal{K}$ can be sampled uniformly and densely. However, based on the depth-defocus blur degree relation

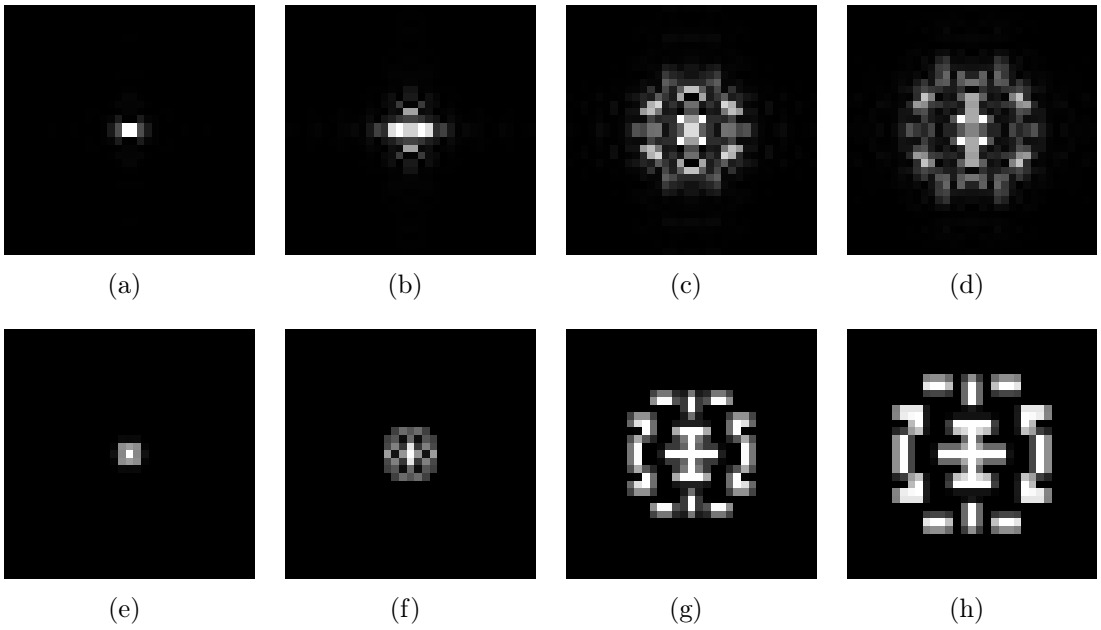
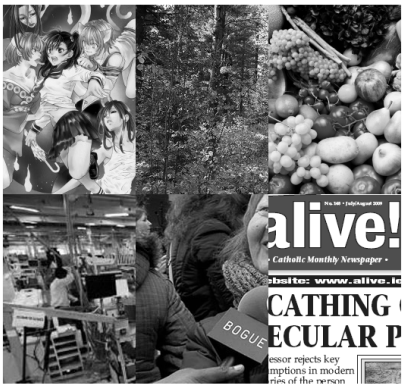


Figure 6.3 *Examples of calculated PSFs for Levin's mask case. (a)-(d) The PSFs calculated based on the wave optics. (e)-(h) The PSFs calculated based on the geometrical optics, with the same camera settings and at the same depths.*

described in Figure 2.8, we can infer that for a fixed blur discrimination ability, the depth resolution provided by the defocus blur cue decreases as the depth increases. Thus, it is more reasonable to sample PSFs according to the blur discrimination criterion. When the criterion is that two consecutive discrete PSFs must differ at least one pixel in scale, it leads to a $\mathcal{K}$ containing non-uniform depths, where depths can be found by calculating Eq. (2.2) with desired PSF scales $N_{pix}s_{pix}$, where N_{pix} is the number of pixels and s_{pix} is the pixel pitch, which means the physical size of a pixel. Specially, when the camera focuses at the infinity i.e. far away from the camera, we have

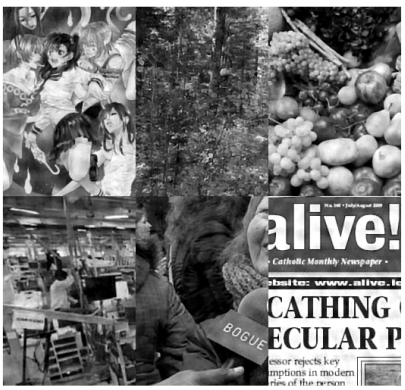
$$d = \frac{fd_L}{N_{pix}s_{pix}}, \tag{6.18}$$

From Eq. (6.18), we can infer that for a fixed amount of PSF scales change, a smaller s_{pix} can lead to a finer depth variance. This observation suggests that using smaller s_{pix} can achieve a higher depth resolution under the same blur discrimination criterion. However, Eq. (2.2) gives a good depth set $\mathcal{K}$ if and only if a sufficiently accurate equivalent thin-lens camera model is available. When this requirement is unsatisfied, the depth set $\mathcal{K}$ can be obtained by uniformly sampling the depth range, yet with a large interval to meet the blur discrimination criterion. The length of this interval may be estimated by using Eq. (2.2). In addition, when a symmetrical

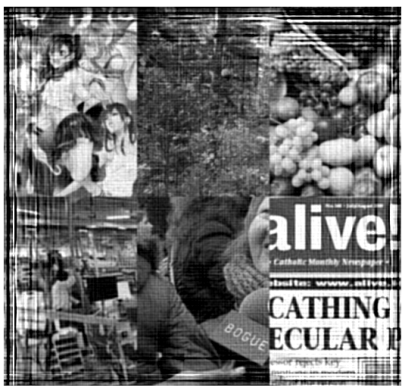


(a) Sharp image.

(b) Defocused Figure 6.4(a).



(c) Restored image provided by Levin's algorithm.



(d) Restored image provided by Zhou's algorithm.

Figure 6.4 *Simple simulation.*

mask is used, it is required to set focus such that the whole scene is on one side of it, e.g. focusing in front of the scene, to avoid the sign problem, as mentioned in Section 5.1.

6.2 Simulations

In this section, three implemented algorithms are tested with images 'captured' by a virtual coded aperture camera. The testing contains two stages using different simulation environments.

In the first stage, the aim is to verify the correctness of algorithms' implementations. Thus, the virtual scene $\boldsymbol{f}_{\boldsymbol{M}}^{0}$ is constructed to be a simple fronto-parallel plane, whose texture is a combination of multiple natural images with different types of contents,

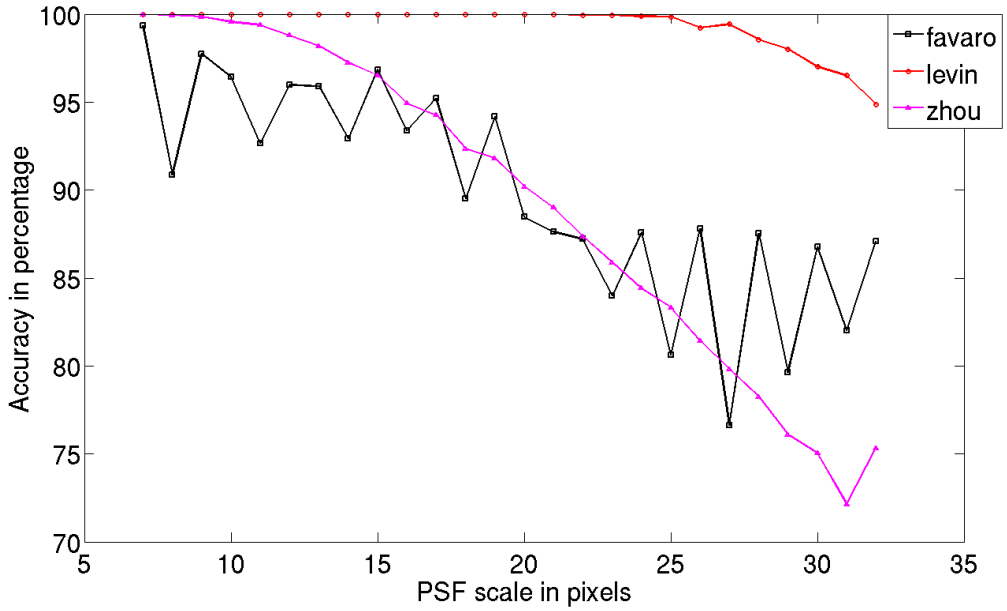


Figure 6.5 *Illustration of testing results.*

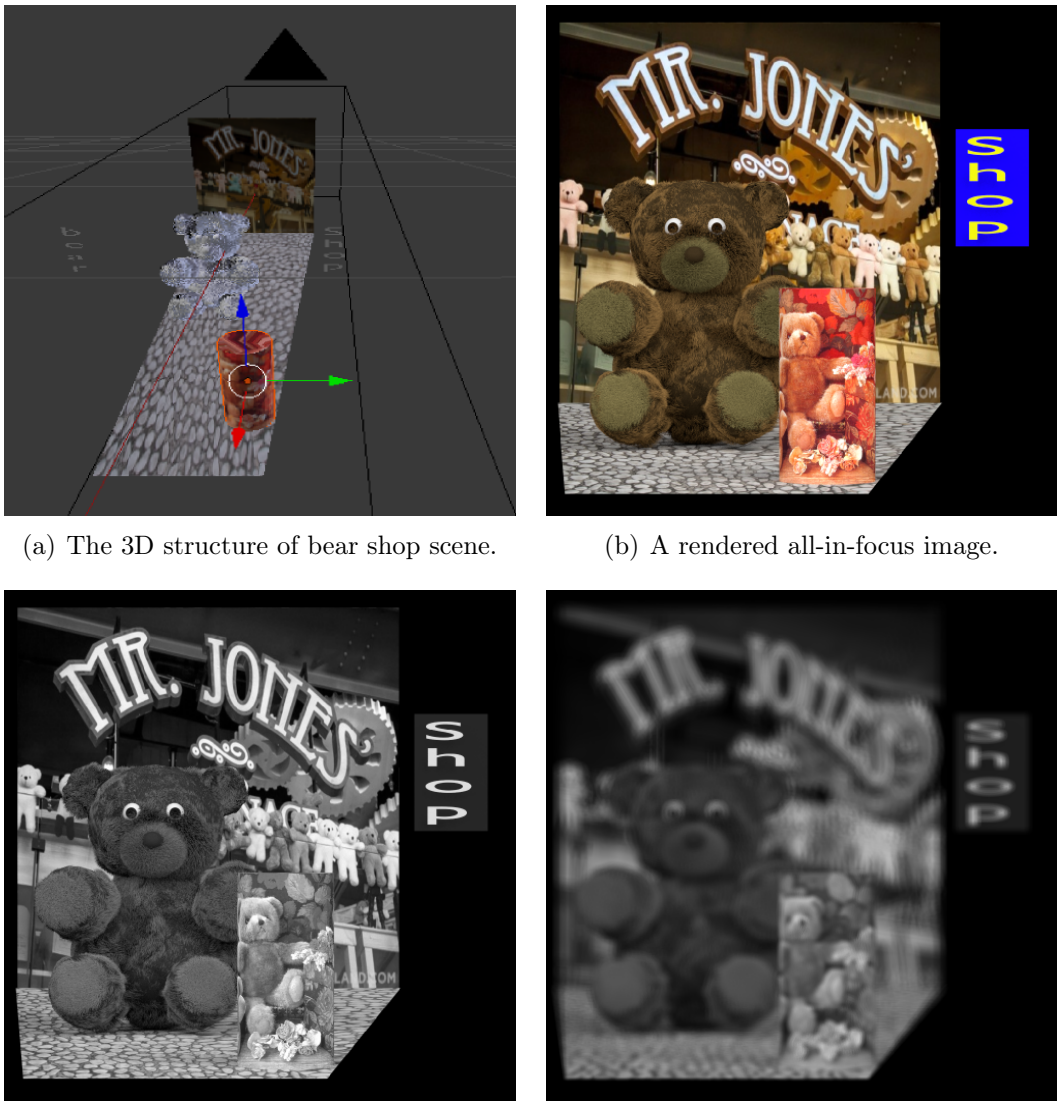
Table 6.4 *The virtual camera settings.*

Item	Value (mm)
aperture size :	16
focal length :	35
focused distance :	1500
pixel pitch :	0.006

as shown in Figure 6.4(a). Then a virtual coded aperture camera whose physical parameters are summarised in Table 6.4 is used to 'capture' defocused images of this scene. In order to eliminate the errors coming from the imperfection of the model of camera imaging system and sampled PSFs, the defocused scene is generated by a simple convolution $\boldsymbol{g_M} = \boldsymbol{f}^0_{\boldsymbol{M}} \otimes \boldsymbol{h}^d + \boldsymbol{\omega_M}$, where d is known to be inside of the depth set $\mathcal{K}$, and $\boldsymbol{\omega_M}$ is additive white Gaussian noise with mean 0 and variance 0.005. In our case, the virtual camera focuses at 1.5 metres, and 26 images with different defocus blur degrees of the plane are 'captured' by putting it at 26 known depths, corresponding to PSF scales from 7 to 32 pixels. For cases of testing Levin's algorithm and Favaro's algorithm, the virtual camera is equipped with Levin's mask; while for testing Zhou's algorithm, Zhou's mask pair is used in turn to acquire a pair of images. An example image 'captured' with Levin's mask at depth corresponding to the PSF scale of 32 pixels is shown in Figure 6.4(b). Please notice that those masks are selected for their demonstrated performance. The results are summarised in Figure 6.5, which shows all three algorithms are

well implemented, and the performance of all three algorithms decrease when the PSF scale becomes larger, which suggests that all three algorithms may have limited working range. Specifically, we notice that in this well controlled simulation environment, Levin's algorithm provides superb results, which is considerably better than results provided by Zhou's algorithm. Since both algorithms follow the same strategy, we interpret this performance difference as a consequence of using different image restoration methods, since under restoration-based strategy, the quality of depth estimation depends on the quality of image restoration and vice versa, as pointed out in Section 4.2. As mentioned in Section 4.2, image restoration in Levin's algorithm is done in the spatial domain by solving Eq. (4.11) via a IRLS algorithm, which does not involve any inverse operation, e.g. division. The restored image of Figure 6.4(b) is shown in Figure 6.4(c). While in Zhou's algorithm, as shown in Eq. (4.16), image restoration is done in the frequency domain via a (generalised) Wiener filter, which involves DFT. The restored image from defocused images 'captured' with Zhou's pair at depth corresponding to the PSF scale 32 pixels is shown in Figure 6.4(d). Unlike Figure 6.4(c), Figure 6.4(d) suffer from ringing artefacts near image boundaries, where the depth estimation fails. These ringing artefacts are caused by DFT, which views the image as a periodic signal in both the spatial and frequency domains. However, as a truncated recording of the scene, an image is rarely periodic. Thus, when the left and right (or top and bottom) sides of images have different values, leakage frequencies will be created. During the deconvolution process, those leakage frequencies near the zero-crossings of the system OTF are amplified and cause the ringing artefacts [58], [25]. This notification suggests that the basic (generalised) Wiener filter used in Zhou's algorithm should be modified by e.g. windowing techniques [25], or we have to keep images having the same values at corresponding boundaries, as we shall do in other simulations below. For Favaro's algorithm, the curve is in a zigzag shape, which indicates that subspaces defined by PSFs are slightly overlapping, and the main reason of this overlapping is that PSFs are not that distinguishable. Apart from this reason, the determination of the rank of subspaces is also important and affects the results heavily. Unfortunately, currently the rank have to be determined based on experience since no reliable methods have been reported, and this is a drawback of Favaro's algorithm. However, once the rank is determined, it will not change since it is independent from images.

In the second stage, a more realistic simulation environment is built for testing the performance of algorithms. Unlike the simple fronto-parallel plane scene used in the first stage, a real 3D scene usually contains objects of complicated surfaces, and their textures may not always be rich. Therefore, as a 3D modelling software, Blender [1] is employed and a 'bear-shop' scene is designed with it. As shown in



(a) The 3D structure of bear shop scene.

(b) A rendered all-in-focus image.

(c) The green channel of Figure 6.6(b).

(d) The green channel of defocused image, captured with Levin's mask.

Figure 6.6 *Illustration of the bear shop scene.*

Figure 6.6(a) and Figure 6.6(b), this scene contains four parts: a cylinder, a bear, a background and a ground, set at different depths. The aperture superposition principle is employed to simulate defocused images 'captured' by coded aperture cameras. As mentioned in Section 3.3, the image captured by a camera with an arbitrary aperture mask pattern can be well approximated by a superposition of images captured with elementary apertures. Since all three masks involved in this simulation are designed by using brute force search as introduced in Section 5.3, it is natural to use $n \times n$ small squares as elementary apertures, where $n = 13$ for Levin's mask and $n = 33$ for Zhou's mask pair. Each elementary square aperture is further divided into finer squares, e.g. $k \times k$ squares, whose size are small enough

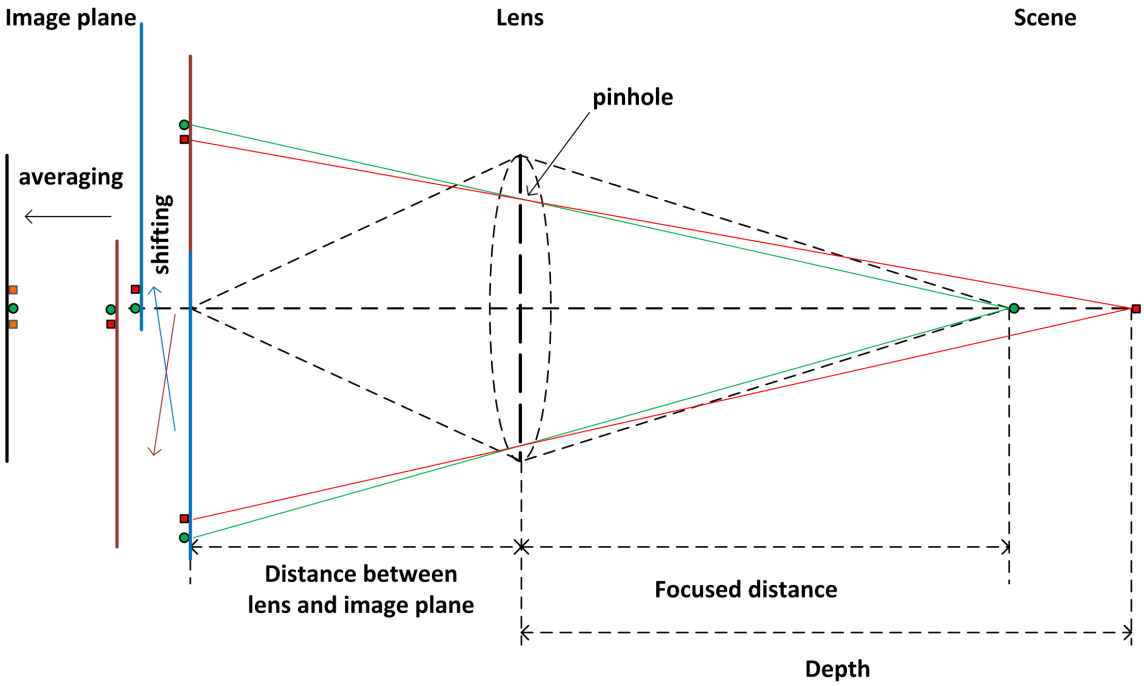


Figure 6.7 *Illustration of shifting and averaging procedure for 1D case.*

to be viewed as 'pinhole's.

Using the thin-lens model, the whole process can be summarised as follows: Firstly, calculating the distance between lens and sensor plane according to the focused distance, as

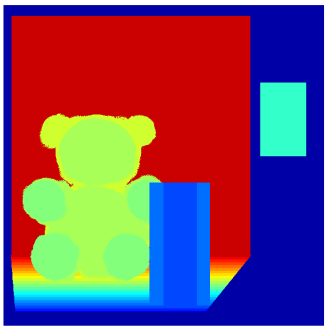
$$l_f = \left(\frac{1}{f} - \frac{1}{d_f}\right)^{-1}. \tag{6.19}$$

This distance l_f is used as the focal length of each 'pinhole' camera.

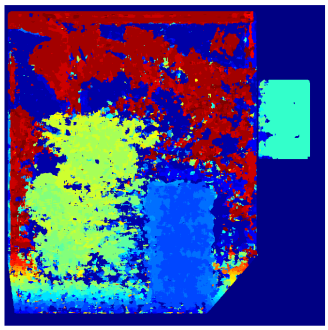
Secondly, the aperture is divided into $m \times m$ small squares where $m = k \times n$, and they are viewed as m^2 'pinhole' apertures. For each 'pinhole' aperture belonging to an opening elementary aperture, an all-in-focus image is rendered according to the pinhole camera model. The lens focusing effect is simulated by shifting the all-in-focus image, and the shifting amount Z is calculated according to Eq. (2.1), as

$$Z = \frac{l_f B}{d_f}, \tag{6.20}$$

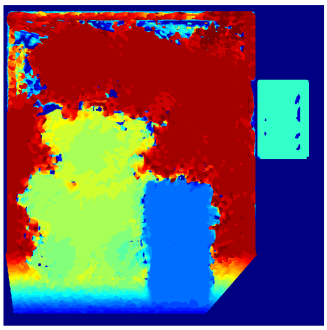
where the baseline B is set as the distance between the position of the 'pinhole' of interest and the aperture centre. The average of all those shifted all-in-focus images is considered as the defocused image. A 1D example is illustrated in Figure 6.7.



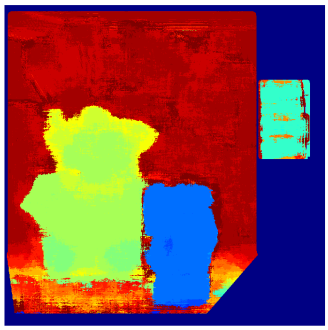
(a) The ground truth depth map in PSF scales.



(b) The result produced by Levin's algorithm.



(c) The result produced by Zhou's algorithm.



(d) The result produced by Favaro's algorithm.

Figure 6.8 *The bear shop scene results.*

Table 6.5 *The noise effect.*

Algorithm \ SNR	Inf	60	50	40	30	20
Zhou	86%	86%	86%	84%	73%	46%
Favaro	82%	82%	80%	74%	52%	23%

In our case, the camera system is again set according to Table 6.4, and the scene depth range is 1.74-2.87 metres. As an example, the green channel of a simulated defocus image with Levin' mask is shown in Figure 6.6(d). Also, the green channel of the all-in-focus image is shown in Figure 6.6(c) as a comparison. Regarding PSFs, since defocused images are rendered based on the geometrical optics, PSFs are also calculated using the geometrical optics based method, at 26 different depths covering the depth range of the 'bear-shop' scene. Three estimated depth maps using three algorithms are shown in Figure 6.8, together with the ground truth depth map. Being restoration-based methods, Levin's algorithm and Zhou's algorithm fail on areas with poor texture, where the image restoration cannot be done properly, especially when only a single image is used like in the Levin's case. However, since

in restoration-based methods, whole images are used for depth estimation on each patches, Levin's algorithm and Zhou's algorithm produce much better results on the ground than Favaro's algorithm, which uses only an image patch to do depth estimation on that patch. On the other hand, Favaro's algorithm is less affected by the poor texture since image restoration is avoided. Please notice that all depth maps shown in Figure 6.8 are raw depth maps without post-processing, and their qualities can be improved by using e.g. MRF as mentioned in Section 4.4.

So far the influence of noise has not been considered. In order to understand the influence of noise, 6 levels of signal-to-noise ratio (SNR) are considered, including $[\text{Inf}, 60, 50, 40, 30, 20]$dB, where Inf means no noise [21]. The performances of Zhou's algorithms and Favaro's algorithm under those SNRs are tested with the 'bear-shop' scene, and the accuracies are summarised in Table 6.5, where the accuracy percentage is calculated by comparing the result to the ground truth depth map, and if the difference is less than or equal to one scale, we accept it as correct. The results show that all tested algorithms can tolerate noises that can be seen in most of the practical cases.

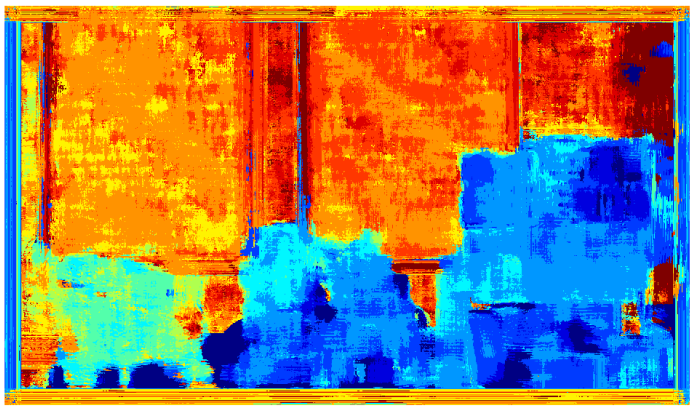
6.3 Experiments

The implemented Favaro's algorithm is tested in a real situation. The real scene has been arranged in a similar way to the 'bear-shop' scene, as shown in Figure 6.9(b). Then the Levin's mask is inserted in a Nikon D5200 DSLR camera mounted with a Nikon 35mm lens, as shown in Figure 6.9(a), and the camera is put in front of the scene such that the depth range is about 2.0-2.5 metres and the focused distance is set at 1.5 metres away from the camera. Coded aperture images are captured under strong white light illumination with ISO 100, to reduce the exposure time and keep sensor noise minimal. In order to minimise the influence of lens distortion, which is not considered during developing algorithms, only the middle areas of captured images are kept. The green channel of the image is used for testing, as shown in Figure 6.9(b). PSFs are calculated at depth range from 1.92-2.7 meters for every 7 cm by using wave optics based method given in Eq. (6.17) for green light corresponding to the green channel of RGB image.

The resulting raw depth maps obtained by using Favaro's algorithm is shown in Figure 6.9(c). We can see that the depths of all objects are approximately obtained. Especially, the upper right corner area, whose depth is further than the maximum depth in the depth set $\mathcal{K}$, is labelled with the maximum PSF scale as expected. However, it is obvious that the result is not that good as in the simulation case. There are several error sources degrading the experimental results. We believe that

(a) The coded aperture camera.

(b) The green channel of the captured defocused image (cropped middle part).



(c) The result produced by Favaro's algorithm.

Figure 6.9 *The real experiment.*

it is mainly due to deviations from the assumptions made in the wave optics based PSF calculation, e.g. aberration-free lens, having an equivalent thin lens model of the camera, etc. It is also worth mentioning the camera noise and measurement errors during the experiment as other sources.

7. CODED APERTURE STEREO CAMERAS

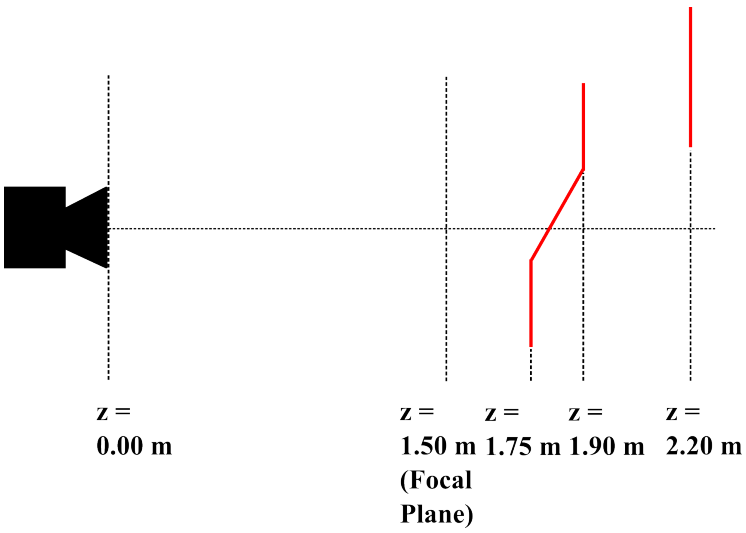
In this chapter, we investigate possible improvements in depth estimation that can be achieved by using stereo cameras with masks, which can be referred as coded aperture stereo cameras. The motivation behind this exploration is twofold. One is to have an integrated system where both the defocus blur and disparity cue are available, since they are considered to be able to provide complementary information in some situations, as mentioned in Section 2.4. The other is to have a single shot multiple coded aperture system for the cases that different masks, e.g. Zhou's mask pair, can be employed simultaneously, since capturing multiple images with different masks from a single view is practically difficult.

7.1 Integrated system

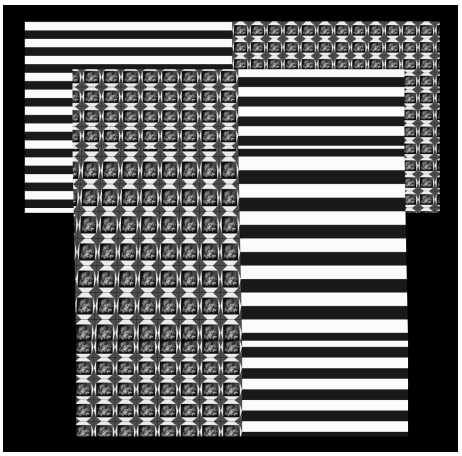
In this section, we aim to develop an integrated system where both the defocus blur cue and the disparity cue are available so that coded aperture and stereo vision based methods, e.g. stereo matching, can work synchronously.

Regarding designing integrated systems, we investigate two questions. One is whether equipping with masks seriously affects the performance of the ordinary stereo matching, which utilise the disparity cue; the other is whether coded aperture can provide useful information in situations where the stereo matching fails. That is, is it worth introducing masks into the system [55]?

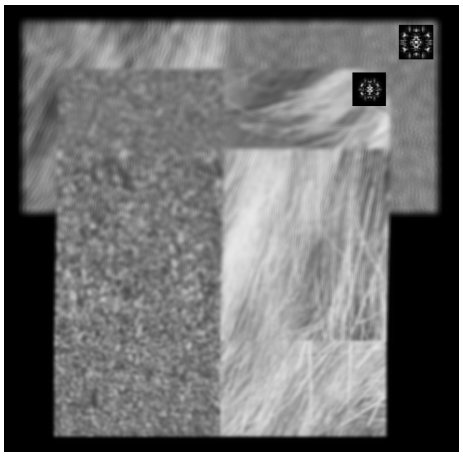
In order to answer aforementioned two questions, a 3D scene denoted as the 'slant' is built in the simulation environment. As shown in Figure 7.1(a), the scene contains three fronto-parallel planes and two of them are connected with a slanted plane. For textures, two cases are considered, one contains repetitive patterns and strips, which both are known to be problematic for stereo matching; the other uses gravel and rabbit's fur as texture, which are good texture for stereo matching in the sense of randomness. Two stereo cameras are assumed to be identical having 35mm lens and focused on 1.5 metres, and the baseline is set to be 5cm. A virtual camera is put in the middle of the baseline, and a middle view image is 'captured' according to Eq. (6.14) with $\lambda = 534nm$. The stereo image pair is generated by shifting



(a) The arrangement of the 'slant' scene.



(b) A left view image captured with pinhole aperture for the problematic texture case.



(c) A right view image captured with Levin's mask for the good texture case, and two example PSFs (scaled by a factor of 3 for visualisation) at depth $d = 1.9m$ and $d = 2.2m$ are shown as well.

Figure 7.1 *Illustration of the simulation environment of the 'slant' scene [55]. Reprinted by permission. ©2014 IEEE.*

the middle view image, and the shifting amount is calculated according to Eq. (2.1). As examples, an image from the left view in the problematic texture case, 'captured' with the ideal pinhole aperture, and an image from the right view in the good texture case, 'captured' with the Levin's mask, are shown in Figure 7.1(b) and Figure 7.1(c), respectively.

To observe whether the performance of stereo matching is seriously affected from equipping the cameras with masks, the same stereo matching algorithm [2] is applied to stereo image pairs 'captured' by stereo cameras with different sets of mask pairs

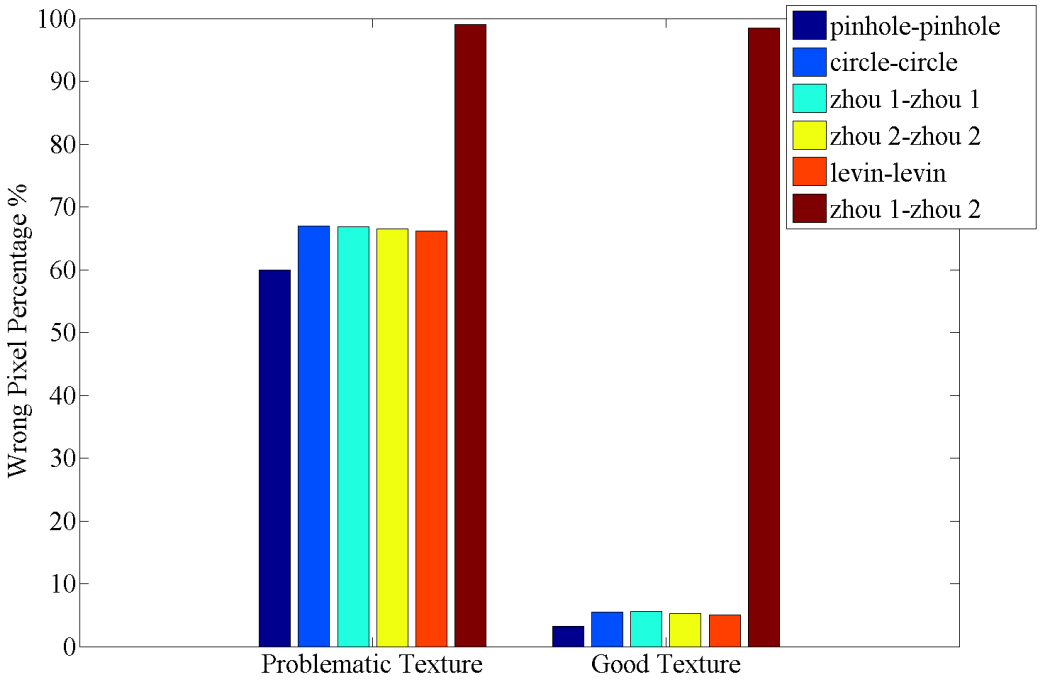
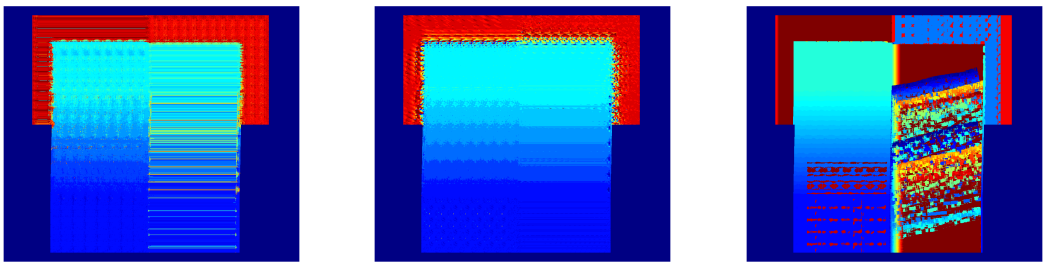


Figure 7.2 *The error percentage of stereo matching for different aperture masks, for both the problematic texture case and the good texture case [55]. Reprinted by permission. ©2014 IEEE.*

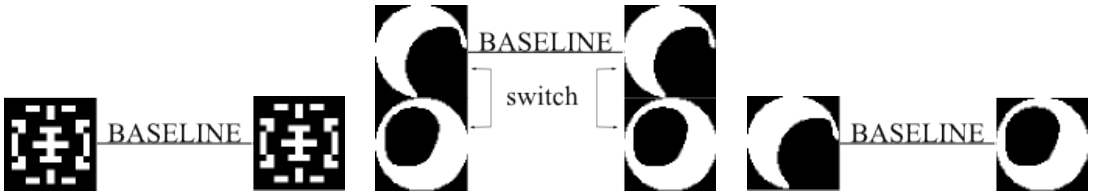


(a) The depth map in PSF scales produced by Favaro's algorithm.

(b) The depth map in PSF scales produced by Zhou's algorithm.

(c) The depth map in disparity values produced by stereo matching [2].

Figure 7.3 *Results produced by three algorithms for the problematic texture case (adapted from Figure 4 in [55]). Reprinted by permission. ©2014 IEEE.*



(a) Stereo cameras with Levin's mask. (b) Two stereo cameras with Zhou's mask pair. (c) Stereo cameras with Zhou's mask pair.

Figure 7.4 *Three proposed camera systems [55]. Reprinted by permission. ©2014 IEEE.*

including the same mask, which are pinhole, circular mask, Levin's mask and Zhou's mask pair (one at a time), in both the problematic texture case and the good texture case. The resulting depth maps are compared with the ground truth depth map, and the accuracy are shown in Figure 7.2. From the results shown in Figure 7.2, we can infer that when two identical masks are used, the influence on the performance of stereo matching is not severe [55], and our observation here is consistent to the human vision case as mentioned in Section 2.4. To answer the second question, Levin's mask and Zhou's mask pair are used from a single view, e.g. the right view, respectively, on the problematic texture case where stereo matching fails. Results obtained by using Favaro's algorithm and Zhou's algorithm, together with the result obtained by stereo matching in pinhole aperture case, are given in Figure 7.3. These results show that for the problematic texture case, coded aperture using the defocus blur cue can give more reliable depth information than stereo matching using the disparity cue, the depth resolution provided by the defocus blur cue is worse than that provided by the disparity cue, though. Similar results are reported by Takeda et al. [51], who notice that on the problematic texture areas, utilising the defocus blur cue can lead to depth map of better quality over the one obtained utilising the disparity cue. Those consistent results are encouraging since they indicate that coded aperture and stereo matching are complementary, in the sense that the former can give more reliable depth information on the problematic texture areas while the latter offers better depth resolution when it works.

Based on the results given above, we proposed two integrated systems as shown in Figure 7.4(a) and Figure 7.4(b). In the first system, two cameras are both equipped with Levin's mask; while in the second system, two more cameras are employed, so that in both views we have a pair of images captured with Zhou's mask pair. In both systems, coded aperture and stereo matching can both work independently with minimal influences on each other, and thus it can produce both a depth map in disparity values and a depth map in PSF scales. When two depth maps contains complementary information, they can be merged by using e.g. MRF [52] to improve

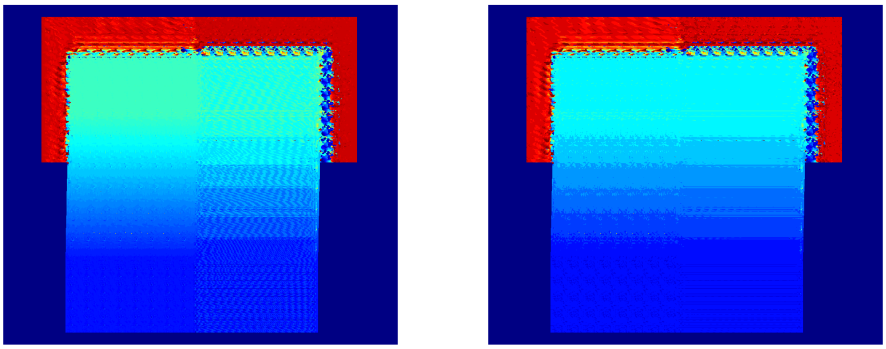
the quality of the depth map.

7.2 Single shot multiple coded aperture system

In this section, we propose a single shot multiple coded apertures system, and a two masks case and the corresponding algorithm are introduced as an example.

Multiple coded apertures systems are of interests for three reasons: Firstly, when only a single image is available, it is impossible to distinguish between focused low texture and blurred high texture, and this ambiguity is a result of losing information, as mentioned in Section 4.1. However, if multiple images captured with different masks of the same scene are available, this ambiguity can be resolved since each of those images may contain different information that can be used as a compensation for other images. This compensation is especially strong when those images are captured with complementary masks e.g. Zhou's mask pair. Secondly, as mentioned in Section 5.3, a single mask can hardly have desired properties for both depth estimation and image restoration simultaneously since they are contradictory, while when multiple masks are available, desired properties for both problems can be satisfied at the same time, e.g. with Zhou's mask pair. Thirdly, according to the analyses and results given in Chapter 5, a desired single mask for depth estimation should be of a symmetrical pattern, which means that it suffers from the sign problem mentioned in Section 5.1. Consequently, depth estimation can only be done on one side of the focused distance. However, this sign problem can be easily avoided by using multiple masks, e.g. Zhou's mask pair where two masks both are of asymmetric patterns, and thus the depth range can be largely extended by focusing at the middle of the scene. Those reasons show the benefits to use multiple coded apertures, and thus form a solid motivation to develop and use multiple masks systems.

Typically, when multiple masks are used, it is required that multiple images are captured with different masks from the same view to guarantee that images are well aligned, which is fundamental for DfD, as pointed out in Section 4.1. In order to satisfy this requirement, several methods for capturing multiple images have been reported. One method is to manually switch lenses of different masks during capturing, and the misalignment introduced during switching lenses is corrected by using affine transformation afterwards [59]. To avoid switching lenses, a pattern scroll or a liquid crystal array (LCA) [24] or a liquid crystal on silicon (LCoS) [35] can be employed to make a programmable aperture camera whose aperture mask can be dynamically changed. Also, a beam splitter can be employed to create two identical views for different masks. However, for using those methods, either an user should be present or complicated modifications/equipments are required. Facing this



(a) Depth map in disparity values. (b) Depth map in PSF scales.

Figure 7.5 *The results produced by the proposed algorithm on the 'slant' scene for the problematic texture case (adapted from Figure 5 in [55]). Reprinted by permission. ©2014 IEEE.*

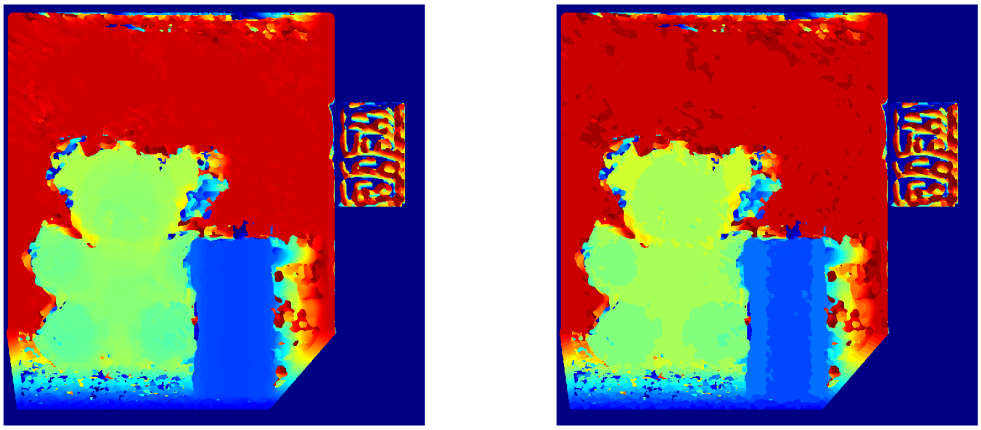
problem, we instead propose a multi-view coded aperture system where each camera is equipped with a mask. For the case of using Zhou's mask pair, it becomes a coded aperture stereo system, as shown in Figure 7.4(c).

Compared to aforementioned other methods, the proposed system has minimal modification on the lens and does not require user manipulation. However, the requirement set by coded aperture is violated, since two images are captured from different views. This violation can be solved by processing captured images. Intuitively, for pixels of a particular depth, misalignment of them in two views can be corrected if two images are shifted by the correct disparity value. Then for those aligned pixels, the requirement is satisfied and thus DfD algorithms, e.g. Zhou's algorithm, should be able to be applied. This can be done for all possible depths and thus all pixels are covered. As shown in Figure 2.7, there exists a linear relation between the defocus blur cue and the disparity cue, which suggests an one-to-one mapping between the disparity value and the PSF scale. However, in most practical cases the depth resolution achieved by coded aperture is coarser than the one achieved by stereo matching, e.g. coded aperture usually only work on a set of pre-sampled depths. Due to this resolution mismatch, we instead set a multi-to-one relation between disparity values and a PSF.

Theoretically, the correct disparity-PSF pair will produce the minimum error [51], [55].

Table 7.1 *The stereo version of Zhou's algorithm (adapted from Table 1 in [55]). Reprinted by permission. ©2014 IEEE.*

INPUTS:	
$\left(\boldsymbol{g}_{\boldsymbol{M}_L}, \boldsymbol{g}_{\boldsymbol{M}_R}\right)$:	captured left and right images;
PSFs :	PSF pairs pre-sampled at a set of depth $\mathcal{K}$;
	each pair is denoted as $\left(\boldsymbol{h}_L^{\boldsymbol{c}_1,d_k}, \boldsymbol{h}_R^{\boldsymbol{c}_2,d_k}\right)$;
STEPS:	
1 :	For each PSFs pair $\left(\boldsymbol{h}_L^{\boldsymbol{c}_1,d_k}, \boldsymbol{h}_R^{\boldsymbol{c}_2,d_k}\right)$ at depth d_k
2 :	Find the associated disparity range S_k of d_k;
3 :	For each $disp$ in S_k
4 :	$\boldsymbol{g}_{\boldsymbol{M}_L}^{prime} = \boldsymbol{g}_{\boldsymbol{M}_L}(x - disp, y)$
5 :	Obtain $\hat{\boldsymbol{F}}_{\boldsymbol{M}}^{k}$ by using Eq. (4.16)
6 :	End for
7 :	End for
8 :	Obtain depth maps by solving Eq. (7.1), $\forall pixel$



(a) The depth map in disparity values. (b) The depth map in PSF scales.

Figure 7.6 *The results produced by stereo version of Zhou's algorithm on the bear shop scene.*

Thus, we can modify Eq. (4.17) to be

$$\boldsymbol{D_M}^*[l], \boldsymbol{Disp_M}^*[l], \boldsymbol{f_M}^*[l] = \underset{d_k, disp, \hat{\boldsymbol{f}}_{\boldsymbol{M}}^{k}}{\arg\min} \sum_{\boldsymbol{L}} \sum_{n=1}^{N} \left(|\boldsymbol{g}_{\boldsymbol{M}_n}(disp) - \boldsymbol{h}^{\boldsymbol{c}_n,d_k} \otimes \hat{\boldsymbol{f}}_{\boldsymbol{M}}^{k}|_2^2 \right). \tag{7.1}$$

The stereo version of Zhou's algorithm according to the analysis above is summarised in Table 7.1.

The proposed coded aperture stereo system utilising Zhou's mask pair and modified Zhou's algorithm are tested with two different simulated scenes. The first scene is the 'slant' scene described in Section 7.1 with problematic texture, and resulting depth map in disparity values and depth map in PSF scales are shown in Figure 7.5. Comparing Figure 7.3(b) and Figure 7.5(b), we can say that the stereo version of Zhou's algorithm produces as good depth map in PSF scales as the original Zhou's algorithm. Moreover, it simultaneously provides a depth map in disparity values, as shown in Figure 7.5(a), which has significantly increased quality compared to the one obtained by directly applying stereo matching on images, shown in Figure 7.3(c). However, it is worth pointing out that this depth map in disparity values is still in the depth resolution provided by the defocus blur cue, since from the stereo matching point of view, Zhou's algorithm in fact is used as a criterion for evaluating stereo correspondence, and this criterion is too coarse to reach the disparity resolution. On the other hand, using stereo cameras unavoidably amplifies the occlusion problem, which is less important in the single view case. This occlusion problem is more visible in the second test, where the 'bear-shop' scene described in Section 6.3 is employed, and the baseline between two cameras are set to be 10 cm. The resulting depth map in disparity values and depth map in PSF scales are shown in Figure 7.6. Interestingly, by jointly using two cues, we can see that the depth map in PSF scales shown in Figure 7.6(b) is in fact better than that is shown in Figure 6.8(c), except suffering from heavy occlusions.

8. DISCUSSION AND CONCLUSION

This thesis studies the problem of depth from defocus from 2D images captured by cameras equipped with coded masks. Mainly two cases are considered: one is analysing the coded aperture for depth estimation in a single view; the other is exploring the possibility of combining coded aperture and stereo vision based methods e.g. stereo matching.

In the first part, analyses on the single view coded aperture technique show that this technique has deficiencies in three aspects which limit its applications. The first aspect is the defocus blur cue it utilises, and the main deficiency is that the defocus blur cue is too vague. In both human vision and computer vision studies, it has been found that the depth-defocus blur degree relation is rather similar to the depth-disparity relation, apart from a scale. In computer vision, those two relations are shown to have the same form, and the lens aperture diameter in the monocular vision serves the role of the baseline in the stereo vision. However, since in most of the practical cases the lens aperture diameter is considerably shorter than the baseline, for the same amount of depth variance, the variance of defocus blur degree is much less significant than the disparity value variance. Due to this scale difference, the depth resolution provided by the defocus blur cue is much less than the one given by the disparity cue. Therefore, practically the defocus blur cue is suggested as a qualitative depth cue, and when it is used as the main depth cue, only a depth information with coarse resolution can be expected.

The second aspect is extracting the depth blur cue encoded in images, and the main deficiency is that extracting the defocus blur cue is an ill-posed problem, whose solution is considerably hard to acquire. Regarding the algorithms to extract the defocus blur cue from images for depth estimation, two strategies have been introduced. When the restoration-based strategy is employed, the quality of depth estimation largely depends on the quality of image restoration. However, due to the information loss and noise contamination during the image formation and recording process, the image restoration is a highly ill-posed problem itself. Although additional information can be introduced by e.g. a well chosen image prior, image restoration is still a hard problem. When image restoration is done in the spatial

domain by using, e.g an iterative re-weighted least squares algorithm like in Levin's algorithm, the algorithm might not converge to the global minimum for the cases where the objective function is non-convex (due to image prior), and thus may give less satisfactory results. It is also worth mentioning that those algorithms are usually computationally demanding and time consuming. To achieve image restoration in the frequency domain, care must be taken if discrete Fourier transform (DFT) is employed. Since the captured images are truncated and discretised signals, DFT may introduce ringing artefacts due to discontinuities of boundary values, as shown in Figure 6.4(d), where a generalised Wiener filter is used to restore the image. On the other hand, when the restoration-free strategy is employed, the problem remains ill-posed and can only be solved in areas with sufficiently rich textures. The success of depth estimation is determined by the quality of a subspace projector or a filter bank constructed for each PSF. However, there are practical issues in constructing those subspace projectors or filter banks. For example, in Favaro's algorithm it remains unclear how to determine the rank of a subspace. Regarding obtaining the optimal filter bank, using statistical learning methods like AMA seems promising, but currently they can only be applied on PSFs that are radially symmetrical. The training procedure for learning the subspace projectors is the main part of those approaches. It becomes time consuming and computationally complex when the number and/or size of PSF increases. Furthermore, the procedure needs to be repeated for different scene depth ranges that correspond to different sets of discrete depths at which PSFs are to be calculated. Last but not least, all algorithms considered in this thesis require PSFs (sometimes equivalently a set of blurred images) pre-sampled at a set of depths, since they are the main ingredients of depth from defocus approaches. However, it is difficult to obtain them accurately, either through experimental measurements or mathematical calculations. Experimental ways might provide satisfactory results in most of the cases since it eliminates the difficulty of system modelling. However, they make the approach impractical due to the necessity of repeating the measurement process for each different scene depth range.

The third aspect is the coded mask pattern, and the main deficiency is that masks optimised under certain conditions are not necessarily optimal for other cases. Several optimised masks have been designed according to different criteria for different purposes. Most of those masks have been designed under the assumption of geometrical optics and this limits the search space of mask patterns, since the optimal mask is searched within a coarse resolution signal space to get rid of diffraction effects. Furthermore, those evaluation criteria are usually derived based on the principle of a certain type of DfD algorithms, so the resulting optimised masks may not be the

best choices, if an algorithm from other types is used. In addition, those masks are only optimised for discriminating a few of PSF scales (corresponding to specific scene depth range), under certain camera parameters and settings. Thus, it is unconvincing that those masks are also optimal for other scenarios. Therefore, a more ideal case for optimising mask patterns would be to have a standard procedure that can be applied in different scenarios. Due to the deficiencies in the mentioned three aspects, further studies on coded aperture technique for depth estimation are needed.

In the second part of the thesis, the combinations of stereo vision and coded aperture have been investigated to explore possible improvements that can be achieved in depth estimation. Two types of multiple coded aperture systems have been proposed and tested via simulations. In the first type, where the same mask is employed in both stereo cameras, it has been observed that coded aperture technique and stereo matching can be applied independently without suffering degradation in the performance of usual stereo matching. It has been shown that having such a system, the stereo vision based depth estimation can be complemented with the valuable information obtained by using the coded aperture technique, in the cases where stereo matching suffers from the correspondence problem, e.g. repetitive textures or occlusions. In the second type of single shot multiple coded aperture system, each different mask has been employed in different cameras in a stereo arrangement to get a single shot system which does not require changing the masks. The relation between the disparity cue and the defocus blur cue has been employed to have a modified coded aperture algorithm tailored for the proposed stereo system. The modified algorithm has been demonstrated to be able to provide depth maps in both disparity values and defocus blur degrees simultaneously. Moreover, it has been shown that by using the proposed method, valuable results can be obtained even in the problematic cases for the standard stereo matching, e.g. repetitive texture, edges along the epipolar lines. All those observations demonstrate that coded aperture technique can serve as a complementary approach to stereo vision, if the defocus blur cue can be correctly extracted.

In conclusion, although it is hard to suggest the coded aperture technique itself as a primary choice for depth estimation, due to the deficiencies discussed above, it may be considered as a valuable complementary technique to other depth estimation approaches like stereo matching.

BIBLIOGRAPHY

[1] "Blender," Available: http://www.blender.org/about/.

[2] W. Abbeloos, "Real-time stereo vision," Master's thesis, Karel de Grote-Hogeschool University College, May 2010.

[3] M. Aggarwal and N. Ahuja, "A pupil-centric model of image formation," *International Journal of Computer Vision*, vol. 48, no. 3, pp. 195–214, 2002.

[4] M. Bertero and P. Boccacci, *Introduction to Inverse Problems in Imaging*. CRC Press, 1998, 352 p.

[5] A. Blake, P. Kohli, and C. Rother, Eds., *Markov random fields for vision and image processing*. MIT Press, 2011, 472 p.

[6] J. Burge and W. S. Geisler, "Optimal defocus estimation in individual natural images," *Proceedings of the National Academy of Sciences*, vol. 108, no. 40, pp. 16 849–16 854, 2011.

[7] A. Chakrabarti, T. Zickler, and W. Freeman, "Analyzing spatially-varying blur," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, Jun 2010, pp. 2512–2519.

[8] E. R. Dowski and W. T. Cathey, "Single-lens single-image incoherent passive-ranging systems," *Applied Optics*, vol. 33, no. 29, pp. 6762–6773, Oct 1994.

[9] E. R. Dowski and W. T. Cathey, "Extended depth of field through wave-front coding," *Applied Optics*, vol. 34, no. 11, pp. 1859–1866, Apr 1995.

[10] J. Ens and P. Lawrence, "An investigation of methods for determining depth from focus," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 15, no. 2, pp. 97–108, Feb 1993.

[11] H. Farid and E. P. Simoncelli, "Range estimation by optical differentiation," *JOSA A*, vol. 15, no. 7, pp. 1777–1786, 1998.

[12] P. Favaro and S. Soatto, *3-D Shape Estimation and Image Restoration-Exploiting Defocus and Motion-Blur*. Springer-Verlag, 2007, 249 p.

[13] W. S. Geisler, J. Najemnik, and A. D. Ing, "Optimal stimulus encoders for natural tasks," *Journal of vision*, vol. 9, no. 13, pp. 1–16, 2009.

[14] I. Gheta, C. Frese, M. Heizmann, and J. Beyerer, "A new approach for estimating depth by fusing stereo and defocus information," in *GI Jahrestagung (1)'07*, 2007, pp. 26–31.

[15] J. Goodman, *Introduction to Fourier Optics*, 3rd ed. Roberts and Company Publishers, 2004, 491 p.

[16] R. Held, E. Cooper, and M. Banks, "Blur and disparity are complementary cues to depth," *Current Biology*, vol. 22, no. 5, pp. 426 –431, 2012.

[17] S. Hiura and T. Matsuyama, "Depth measurement by the multi-focus camera," in *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, Jun 1998, pp. 953–959.

[18] N. Joshi, R. Szeliski, and D. Kriegman, "Psf estimation using sharp edge prediction," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, Jun 2008, pp. 1–8.

[19] E. Kee, S. Paris, S. Chen, and J. Wang, "Modeling and removing spatially-varying optical blur," in *Computational Photography (ICCP), 2011 IEEE International Conference on*, Apr 2011, pp. 1–8.

[20] D. Lanman, R. Raskar, and G. Taubin, "Modeling and synthesis of aperture effects in cameras," in *Proceedings of the Fourth Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*, ser. Computational Aesthetics'08. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2008, pp. 81–88.

[21] P. Leclercq and J. Morris, "Robustness to noise of stereo matching," in *Image Analysis and Processing, 2003.Proceedings. 12th International Conference on*, Sept 2003, pp. 606–611.

[22] A. Levin and Y. Weiss, "User assisted separation of reflections from a single image using a sparsity prior," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 9, pp. 1647–1654, Sept 2007.

[23] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," in *ACM SIGGRAPH 2007 Papers*, ser. SIGGRAPH '07. New York, NY, USA: ACM, 2007. [Online]. Available: http://doi.acm.org/10.1145/1275808.1276464

[24] C.-K. Liang, T.-H. Lin, B.-Y. Wong, C. Liu, and H. H. Chen, "Programmable aperture photography: Multiplexed light field acquisition," in *ACM SIGGRAPH 2008 Papers*, ser. SIGGRAPH '08. New York, NY, USA: ACM, 2008, pp. 55:1–55:10.

[25] H. Lim, K.-C. Tan, and B. Tan, "Edge errors in inverse and wiener filter restorations of motion-blurred images and their windowing treatment," *CVGIP: Graphical Models and Image Processing*, vol. 53, no. 2, pp. 186 – 195, 1991.

[26] J. Lin, X. Ji, W. Xu, and Q. Dai, "Absolute depth estimation from a single defocused image," *Image Processing, IEEE Transactions on*, vol. 22, no. 11, pp. 4545–4550, Nov 2013.

[27] C. Liu, W. Freeman, R. Szeliski, and S. B. Kang, "Noise estimation from a single image," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, Jun 2006, pp. 901–908.

[28] J. A. Marshall, C. A. Burbeck, D. Ariely, J. P. Rolland, and K. E. Martin, "Occlusion edge blur: a cue to relative visual depth," *Journal of The Optical Society of America A*, vol. 13, no. 4, pp. 681–688, Apr 1996.

[29] M. Martinello and P. Favaro, "Single image blind deconvolution with higher-order texture statistics," in *Video Processing and Computational Video*, ser. Lecture Notes in Computer Science, D. Cremers, M. Magnor, M. Oswald, and L. Zelnik-Manor, Eds. Springer Berlin Heidelberg, 2011, vol. 7082, pp. 124–151.

[30] B. Masia, L. Presa, A. Corrales, and D. Gutierrez, "Perceptually optimized coded apertures for defocus deblurring," *Computer Graphics Forum*, vol. 31, no. 6, pp. 1867–1879, 2012.

[31] G. Mather, "Image blur as a pictorial depth cue," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 263, no. 1367, pp. 169–172, 1996.

[32] G. Mather, "The use of image blur as a depth cue," *Perception*, vol. 26, no. 9, pp. 1147–1158, 1997, pion Ltd, London, www.pion.co.uk and www.envplan.com.

[33] G. Mather and D. R. R. Smith, "Depth cue integration: stereopsis and image blur," *Vision Research*, vol. 40, no. 25, pp. 3501 – 3506, 2000.

[34] G. Mather and D. R. R. Smith, "Blur discrimination and its relation to blur-mediated depth perception," *Perception*, vol. 31, no. 10, pp. 1211–1219, 2002.

[35] H. Nagahara, C. Zhou, T. Watanabe, H. Ishiguro, and S. Nayar, "Programmable aperture camera using lcos," in *Computer Vision - ECCV 2010*, ser. Lecture Notes in Computer Science, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Springer Berlin Heidelberg, 2010, vol. 6316, pp. 337–350.

[36] A. P. Pentland, "A new sense for depth of field," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-9, no. 4, pp. 523–531, Jul 1987.

[37] N. Qian, "Binocular disparity and the perception of depth," *Neuron*, vol. 18, no. 3, pp. 359–368, 1997.

[38] A. Rajagopalan and S. Chaudhuri, "An mrf model-based approach to simultaneous recovery of depth and restoration from defocused images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 7, pp. 577–589, Jul 1999.

[39] A. Rajagopalan, S. Chaudhuri, and U. Mudenagudi, "Depth estimation and image restoration using defocused stereo pairs," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 11, pp. 1521–1525, Nov 2004.

[40] J. C. Read, "Visual perception: Understanding visual cues to depth," *Current Biology*, vol. 22, no. 5, pp. R163 – R165, 2012.

[41] P. R. Sanz, B. R. Mezcua, and J. M. Sánchez Pena, "Depth estimation - an introduction," in *Current Advancements in Stereo Vision*, A. Bhatti, Ed. InTech, 2012, Available: http://www.intechopen.com/books/current-advancements-in-stereo-vision/depth-estimation-an-introduction.

[42] A. Saxena, J. Schulte, and A. Y. Ng, "Depth estimation using monocular and stereo cues," in *IJCAI*, vol. 7, 2007.

[43] Y. Schechner and N. Kiryati, "Depth from defocus vs. stereo: How different really are they?" *International Journal of Computer Vision*, vol. 39, no. 2, pp. 141–162, 2000.

[44] C. M. Schor and I. Wood, "Disparity range for local stereopsis as a function of luminance spatial frequency," *Vision Research*, vol. 23, no. 12, pp. 1649 –1654, 1983.

[45] A. Sellent and P. Favaro, "Which side of the focal plane are you on?" in *Computational Photography (ICCP), 2014 IEEE International Conference on*, May 2014, pp. 1–8.

[46] A. Sellent and P. Favaro, "Optimized aperture shapes for depth estimation," *Pattern Recognition Letters*, vol. 40, no. 0, pp. 96 – 103, 2014.

[47] R. Snowden, P. Thompson, and T. Troscianko, *Basic vision: an introduction to visual perception*, revised ed. Oxford University Press, 2012, 424 p.

[48] E. M. Stein and R. Shakarchi, *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*. Princeton University Press, 2005, 424 p.

[49] R. Szeliski, *Computer vision: algorithms and applications*. Springer, 2010, 812 p.

[50] Y. Takeda, S. Hiura, and K. Sato, "Coded aperture stereo: For extension of depth of field and refocusing," in *VISAPP 2012 - Proceedings of the International Conference on Computer Vision Theory and Applications*, vol. 1, 2012, pp. 103–111.

[51] Y. Takeda, S. Hiura, and K. Sato, "Fusing depth from defocus and stereo with coded apertures," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, Jun 2013, pp. 209–216.

[52] M. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, Dec 2013, pp. 673–680.

[53] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, "Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing," in *ACM SIGGRAPH 2007 Papers*, ser. SIGGRAPH '07. New York, NY, USA: ACM, 2007.

[54] D. Vishwanath, "The utility of defocus blur in binocular depth perception," *i-Perception*, vol. 3, no. 8, pp. 541–546, 2012.

[55] C. Wang, E. Sahin, O. J. Suominen, and A. P. Gotchev, "Depth estimation by combining stereo matching and coded aperture," in *Visual Communications and Image Processing (VCIP), IEEE Conference on*, Dec 2014, pp. 291–294.

[56] A. B. Watson and A. J. Ahumada, "Blur clarified: A review and synthesis of blur discrimination," *Journal of Vision*, vol. 11, no. 5, 2011.

[57] Y. Weiss and W. Freeman, "What makes a good model of natural images?" in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, Jun 2007, pp. 1–8.

[58] J. Woods, J. Biemond, and A. Tekalp, "Boundary value problem in image restoration," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.*, vol. 10, Apr 1985, pp. 692–695.

[59] C. Zhou, S. Lin, and S. Nayar, "Coded aperture pairs for depth from defocus," in *Computer Vision, 2009 IEEE 12th International Conference on*, Sept 2009, pp. 325–332.

[60] C. Zhou, S. Lin, and S. Nayar, "Coded aperture pairs for depth from defocus and defocus deblurring," *International Journal of Computer Vision*, vol. 93, no. 1, pp. 53–72, 2011.

[61] C. Zhou and S. Nayar, "What are good apertures for defocus deblurring?" in *Computational Photography (ICCP), 2009 IEEE International Conference on*, Apr 2009, pp. 1–8.

[62] X. Zhu, S. Cohen, S. Schiller, and P. Milanfar, "Estimating spatially varying defocus blur from a single image," *Image Processing, IEEE Transactions on*, vol. 22, no. 12, pp. 4879–4891, Dec 2013.